COUNTRY AND LANGUAGE LEVEL DIFFERENCES IN MULTILINGUAL DIGITAL LIBRARIES

DISSERTATION

zur Erlangung des akademischen Grades

Doctor philosophiae (Dr. phil.)

im Fach Bibliotheks- und Informationswissenschaft

eingereicht an der

Philosophischen Fakultät I

Humboldt-Universität zu Berlin

von

Maria Luisa Gäde

Präsident der Humboldt-Universität zu Berlin:

Prof. Dr. Jan-Hendrik Olbertz

Dekan der Philosphischen Fakultät I:

Prof. Michael Seadle, PhD

Gutachter:

1. Prof. Vivien Petras, PhD

2. Paul Clough, PhD

Datum der Einreichung: 10.12.2013

Datum der Disputation: 05.02.2014

ABSTRACT

COUNTRY AND LANGUAGE LEVEL DIFFERENCES IN MULTILINGUAL DIGITAL LIBRARIES

by Maria Luisa Gäde

Do digital libraries speak the language of their users? While the importance of multilingual access to information systems is unquestioned, it remains unclear if and to what extent system functionalities, interfaces or interaction patterns need to be adapted according to country or language specific user behaviors. This dissertation postulates that the identification of country and language level differences in user interactions is a crucial step for designing effective multilingual digital libraries. The degree to which digital libraries adapt to them shapes their acceptance within different country groups and language communities.

Due to the lack of comparable studies and analysis approaches, the research in this dissertation identifies indicators that could show differences in the interactions of users from different countries or languages:

RQ1: Which indicators in log files can be leveraged to identify country and language context within multilingual digital libraries?

A customized logging format and logger (Europeana Language Logger) is developed in order to trace these variables in a digital library. As a case study, the dissertation presents the results of a log file analysis of multilingual access to Europeana, the digital library for Europe's cultural institutions such as libraries, audio-visual archives, and museums. In total, 1,071,872 sessions from 21 countries are analyzed with respect to 20 variables and tested for the hypothesized country and language level differences:

RQ2: Does usage data indicate country or language specific interaction patterns?

> $\mathbf{H_0}$: Sessions from different countries and language backgrounds show the same interactions.
>
> $\mathbf{H_1}$: Country and language level differences exist between sessions.

For each investigated variable, differences between country groups are presented and discussed.

To generalize the findings from the case study, the individual variables are prioritized by determining which ones show the most significant country and language level differences:

RQ3: Which variables gathered by log files uncover significant country and language specific differences in user interactions?

Based on a country cluster analysis, 11 out of 20 variables are classified as high impact indicators, having a strong influence on country attributes. Substantial country and language level differences are observed for the usage and preference of the Europeana language interfaces as well as for the refinement and selection of native language content. Country profiles are developed as a tool to visualize different characteristics in comparison.

The methodology and analysis developed in this thesis generate insights for country and language dependent focus points in system design and can also lead to future research dealing with single aspects in more detail. The work concludes with an outlook on future and complementary work in the field of user studies in multilingual environments such as digital library portals, focusing on purposeful correlations, the impact of the interface language change and native content on user interactions.

ZUSAMMENFASSUNG

COUNTRY AND LANGUAGE LEVEL DIFFERENCES IN MULTILINGUAL DIGITAL LIBRARIES

von Maria Luisa Gäde

Sprechen Digitale Bibliotheken die Sprache ihrer Nutzer? Während die Bedeutung von mehrsprachigem Zugang zu Informationssystemen unumstritten ist, bleibt es unklar, ob und in welchem Umfang Systemfunktionalitäten und -oberflächen sowie das Interaktionsdesign an länder- bzw. sprachspezifisches Nutzerverhalten angepasst werden muss und sollte. Die Dissertation legt den Fokus auf die Identifikation von länder- und sprachspezifischen Unterschieden in Interaktionen mit dem Informationssystem als entscheidende Voraussetzung für die Entwicklung von mehrsprachigen Digitalen Bibliotheken. Inwieweit Digitale Bibliotheken sich auf die Bedürfnisse internationaler Nutzer einstellen, wird maßgeblich zu ihrer Akzeptanz und Nutzung beitragen.

Durch den Mangel an vergleichbaren Studien -und Analyseansätzen, identifiziert die Studie zunächst Indikatoren, die auf Unterschiede im Verhalten von Nutzern aus unterschiedlichen Ländern und aus unterschiedlichen Sprachgruppen hinweisen können:

RQ1: Welche Indikatoren aus Logdateien können für die Analyse von länder- und sprachspezifische Interaktionen herangezogen werden?

Basierend auf der Selektion von Indikatoren wurde für die Arbeit ein individuell auf die Problematik von mehrsprachigem Zugang zu Informationssystemen angepasstes Logformat und Analysetool entwickelt, der Europeana Language Logger (ELL). Als Fallstudie dient das Europeana Portal, die Digitale Bibliothek für europäische Kulturinstitutionen wie Bibliotheken, audiovisuelle Archive, Museen und Archiven. Die Analyse umfasst insgesamt 1.071.872 Sessions aus 21 Ländern und untersucht 20 ausgewählte Variablen des Nutzerverhaltens und mögliche Beziehungen zwischen ihnen im Hinblick auf folgende Fragestellung:

RQ2: Zeigen Nutzer aus verschiedenen Ländern unterschiedliche Interaktionsmuster?

> **H_0**: Sessions aus unterschiedlichen Länder- und Sprachgruppen weißen die gleichen Interaktionsmuster auf.
>
> **H_1**: Es bestehen länder- und sprachspezifische Unterschiede zwischen den Sessions.

Für alle Variablen und insbesondere für die Wahl der Oberflächensprache sowie die Präferenz für muttersprachliche Ergebnisse wurden signifikante Unterschiede zwischen den Ländern beobachtet.

Um die Erkenntnisse aus der Fallstudie verallgemeinern können, wurde auf der Basis einer Clusteranalyse eine Gewichtung von starken und schwachen Variablen für die Identifizierung von länder- und sprachspezifischen Unterschieden vorgenommen:

RQ3: Welche Variablen aus Logdateien weisen besonders auf länder- und sprachspezifische Interaktionen hin?

Von den 20 untersuchten Variablen, wurden 11 als starke Indikatoren für die Charakterisierung von länder- bzw.- sprachspezifischen Interaktionen klassifiziert. Auf der Grundlage aller Variablen wurden Länderprofile erstellt und grafisch umgesetzt. Diese eignen sich für die Beschreibung und den Vergleich von länder- und sprachspezifischen Interaktionen innerhalb eines bestimmten Systems.

Die Ergebnisse der Dissertation bestätigen, dass die Entwicklung von mehrsprachigen Digitalen Bibliotheken unter der besonderen Berücksichtigung der Anforderungen von internationalen Nutzern einhergehen sollte. Die Arbeit schließt mit einem Ausblick auf zukünftige und ergänzende Studien im Bezug auf das Nutzerverhalten und Voraussetzungen in mehrsprachigen digitalen Bibliotheken. Insbesondere der Einfluss und die Auswirkungen der Oberflächensprache sowie der vorhandenen muttersprachlichen Inhalte auf das Nutzerverhalten, sollten im Fokus zukünftiger Studien stehen.

ACKNOWLEDGMENTS

„Saber leer es saber andar. Saber escribir es saber ascender.“
(To know how to read is to know how to walk. To know how to write is to know how to move forward).
Jose Marti

First and foremost I want to thank my Doktormutter *Vivien Petras* for teaching me how to write (my own book). Her passion for research as well as her tireless efforts and exceptional support are unique and I am deeply grateful for the last years. My special thanks go to my second advisor *Paul Clough* for his valuable questions and input that enriched my thesis and encouraged me to continue, especially in the last period of this project.

I want to thank *Sjoerd* for listening to my idea and bringing it to life. The ELL would not have been possible without his help. It was a great journey working with him and I am sure it is not over until the last digital library has implemented our logger. It was a pleasure to work with *Dirk* on the very first log analysis enjoying his programming power and patience. I am thankful that he still talks to me. Thanks *uncle Basti* and *Robert* for showing me the joy of statistics. I never knew how much fun things like cluster analysis can bring into your life. I am very thankful for my PhD buddy *Marlies* who taught me Austrian swearing and my fellow sufferer *Juliane*. Her ambition and stamina encouraged me whenever I thought I could not finish my work. I hope our paths will always cross.

Big thanks go to my friends, especially *Benni* for reading these pages and *Uli,* my wonderful “Backfisch” friend. You are blessed! *Martina and Fidel* for opening a new world to me. Furthermore, I would like to thank everybody at *La Casa Buena Vista* who kept asking me the painful question: “When will you hand in your dissertation?”

Mein ganz besonderer Dank gilt meiner großartigen Familie Rolf, Kerstin, Mario und Andreas, die mich bedingungslos ertragen und mein Leben täglich bereichern.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDIX A

ABBREVIATIONS

CACAO	Cross Language Access to Catalogues and Online Libraries
CLEF	Cross Language Evaluation Forum
CLIR	Cross Language Information Retrieval
DL	Digital Library
EDL	European Digital Library
HCI	Human Computer Interaction
IIR	Interactive Information Retrieval
IR	Information Retrieval
IP	Internet Protocol
MLIA	Multilingual Information Access
MLIR	Multilingual Information Retrieval
MT	Machine Translation
TLA	Transaction Log Analysis
TEL	The European Library
SERP	Search Engine Result Page
SLA	Search Log Analysis

VARIABLES USED IN LOG ANALYSIS

GL	Language of External Google Referrer
BL	Browser Locale Language
LC	Interface Language Change
UIL	User Interface Language
D	Duration of Sessions
Q	Unique Queries per Session
EA	External Access Point
BR	Bounce Rate
LG	Login
SS	Search Sessions
BS	Browsing Sessions
BRP	Brief Result Paging
FRP	Full Result Paging
LF	Usage of Language Facet
NLF	Selection of Native Language Facet
CF	Usage of Country Facet
NCF	Selection of Native Country Facet
NLC	Selection of Native Language Collections
NCC	Selection of Native Country Collections
OL	Outlinks to Content Provider

1. INTRODUCTION

1.1 THE IMPACT OF COUNTRY AND LANGUAGE CONTEXT

"The Web must allow equal access to those in different economic and political situations; those who have physical or cognitive disabilities; those of different cultures; and those who use different languages with different characters that read in different directions across a page."
(Berners-Lee, 1999, p. 178)

In the past 20 years, the internet has become a global communication channel. With increasing online interaction, individual user differences are transferred to the Internet. Research emphasizes the impact of the user's background and context as an influencing factor when accessing and interacting with information systems (Ford et al., 2001, Lamb and Kling, 2003). Information seeking behavior has been studied with regard to individual differences such as age (Bilal and Kirby, 2002), gender (Halder et al., 2010) or domain knowledge (Clough and Eleta, 2010).

For multilingual information systems, individual differences with respect to the user's country of origin and language skills could be postulated to have an impact. Country specific stereotypes are present in our daily life. Thinking about Great Britain citizens, a nation of football fans comes to mind while mafia scenarios dominate the view of Italy. David Hasselhoff loving Germans are hard working in contrast to Spanish people who tend to arrive late to appointments. Although these examples reflect prejudices rather than reality, most people believe that societies share characteristics of behavior that distinguish them from others (Deutscher, 2011; Chen, 2013). When it comes to information system design, we can ask the question if country and language level differences also manifest in different users' information seeking behavior.

Are information systems for audiences from different countries or with different languages or multilingual information systems an important research topic? With more multilingual content and more multilingual users joining the digital realm every day, studying country and language differences could be crucial for successful system design. The majority of Internet users are non-English speakers. Table 1.1 illustrates the distribution of Internet users per region. Most users are from Asia and Europe, followed by the English speaking North America.

World Region	Internet Users	% of Internet Users
Asia	1,076,681,059	44.8 %
Europe	518,512,109	21.5 %
North America	273,785,413	11.4 %
Latin America / Caribbean	254,915,745	10.6 %
Africa	167,335,676	7.0 %
Middle East	90,000,455	3.7 %
Oceana / Australia	24,287,919	1.0 %
Total	7,017,846,922	100.0 %

Table 1.1 Internet users by geographic region (in millions) (2012)[1]

While Internet users have long reached a multilingual equilibrium, web content is still dominated by one language: English. However, with more non-English users, web content languages also become more varied (Paolillo et al., 2007). Compared to results from early surveys of web site languages with more than 80% of English websites in the nineties[2], today almost 50% non-English content can be observed. The World Wide Web Technology Surveys provides continuous trends of web site languages showing an increase of non-English content (table 1.2).

Language	Websites
English	55.1%
Russian	6.3%
German	5.1%
Spanish, Castilian	4.7%
Chinese	4.4%
French	4.2%
Japanese	4.1%
Portuguese	2.4%
Polish	1.8%
Italian	1.5%

Table 1.2 Top ten languages for websites (2013)[3]

Especially for digitized cultural heritage content, where native languages belong to the cultural context, the importance of multilingual access to digitized content and especially digital heritage is highlighted (UNESCO, 2003a; UNESCO, 2003b). The impact of country and language level differences and their consequences for system design is important in language diverse regions

[1] Internet World Stats - http://www.internetworldstats.com/stats.htm
[2] http://alis.isoc.org/palmares.en.html
[3] Source: http://w3techs.com/technologies/overview/content_language/all

like Europe. The European Union with its currently 24 official languages[4] works towards providing access to information across cultures and languages (Haselhuber, 2012).

Ideally, information systems provide boundless access to information, irrespective of the user's origin and linguistic background (Zhang and Lin, 2007). However, the barrier between the language of a website and the user language is still an open issue. From the user perspective, language skills play an important role when accessing content. For example, monolingual users that do not have knowledge of another language than their native tongue need more comprehensive search assistance when accessing documents in other languages. On the other hand, users with active or passive foreign language skills might be able to inspect non-native content with the help of machine translation options (Peters et al., 2012). However, all user groups want to find results with a single query and do not want or are not able to repeat and translate their information need in different languages.

Search engines like Google[5] already successfully exploit user country information in order to personalize and improve the search experience. For example, Google allows users to either specify their location and preferences or auto-detects the user's location via the IP address or activated location history and redirects him to the appropriate domain. Based on these parameters, search results related to the user's location are presented first in the result list. Different from most other user context information, the location of a user is often transferred to the system when accessing a website. Therefore, this aspect of the user context is particularly suitable for system designers.

Compared to the web and search engines, digital libraries usually serve specific (but global) audiences with content that is often unique, context dependent and difficult to access. In this domain, where content is not available parallel in several languages, overcoming the language barrier becomes even more important to provide universal access. Because of the international audiences and unique, multilingual content, the digital library domain was chosen to study country and language level differences in this research. The following sections introduce dimensions of multilinguality in digital libraries and outline the motivation and structure of the dissertation.

[4] http://ec.europa.eu/languages/languages-of-europe/eu-languages_en.htm
[5] http://www.google.com

1.2 MULTILINGUAL DIGITAL LIBRARIES

With the development of digital libraries, objects can be accessed from users all over the world. Thus, digital libraries face the problem of establishing multilingual access to their collections (Borgman, 1997). Due to the cultural and language diversity of the European Union, especially European applications need to apply multilingual access strategies (Gey et al., 2006).

When researching multilinguality in information systems, three concepts are important to distinguish: multilingual information access (MLIA), multilingual information retrieval (MLIR) and cross-language information retrieval (CLIR). MLIA used as an umbrella term considers all aspects of multilinguality in information systems including accessibility, search, retrieval and inspection of objects regardless of the user or content language. Multilingual information retrieval describes systems that provide multilingual query functionalities and / or content more precisely, whereas cross-lingual information retrieval (CLIR) as part of information retrieval research focuses on the retrieval of documents in other languages than the query language (Oard and Diekema, 1998; Gey et al., 2005). Up to date, most systems provide access to multilingual resources but only support monolingual search functionalities.

Dimensions of multilinguality in digital libraries can be classified according to three perspectives (Oard, 2009):

- User language,
- System language,
- Content language.

The native or preferred languages of users as well as additional language skills influence user needs and requirements. The system language is represented by its interface. Multilingual systems provide localized interface representations for a selected set of languages. Besides the linguistic diversity of users and the respective interfaces, multilingual digital libraries also have to deal with content presented in several languages. The language of content in digital libraries can either be determined on the metadata or the object level. Especially for non-textual objects like images, only metadata information contains language information.

While the technical aspects of multilingual information access (e.g. machine translation) are the focus of much research, fewer studies deal with the user point of view. More recently, interactive information retrieval (IIR) studies put the user-system interaction in the center of analysis (Kelly, 2009). However, only a small proportion of digital library studies focuses on user issues and even less on multilingual or cross-cultural aspects (Liew, 2009). While a lot of

effort has gone into the implementation of multilingual user interfaces, less research has focused on the interaction between the user and the content language (Vassilakaki and Garoufallou, 2013). In her review of studies related to multilinguality in digital libraries, Anne R. Diekema concludes that actual users and their usage of existing multilingual systems need to be the focus of future studies (Diekema, 2012; Chen and Bao, 2009).

When researching country and language level differences of user interactions in digital libraries, the three language dimensions have to be considered. In chapter 5, indicators for country and language level differences based on these perspectives will be defined.

1.3 RESEARCH QUESTIONS AND CONTRIBUTION

This dissertation postulates that the identification of country and language level differences in user interactions is a crucial step for designing effective multilingual digital libraries. The degree to which digital libraries adapt to them shapes their acceptance within different language communities.

So far, language issues in digital libraries have been examined through qualitative studies with a limited number of test users (Agosti et al., 2009b; Marlow et al., 2007; Minelli et al., 2006; Aula and Kellar, 2009). Qualitative research allows identifying individual differences and preferences but cannot determine general patterns. Country level or regional differences were mainly addressed by cultural studies or focused on single aspects like query reformulation patterns (Jesper et al., 2013).

This dissertation proposes a quantitative approach to study country and language level differences through the analysis and interpretation of interactions. An interaction is defined here as the communication between the user and the system under investigation. An in-depth log file analysis was chosen as data collection method as an unobtrusive way to collect and observe usage data from different countries. Interactions represented in log files are understood as traces of user behavior (Jansen, 2009).

A single, but large digital library was used as a case study to analyze. The Europeana portal aggregates content from Europe's national libraries, archives, audio archives and museums. Because of its multilingual content as well as its international audience, Europeana is especially suitable to study country and language level differences in digital libraries. Results derived from aggregated Europeana usage data can also - to a certain extent - be applied to information systems of individual Europeana content providers because both the content and users overlap.

Due to the lack of comparable studies and analysis approaches, the research for this dissertation started by identifying variables that could be assumed to show differences in the interactions of users from different countries or languages (research question 1). A customized logging format and logger was developed in order to trace these variables in a digital library. The collected data from Europeana user interactions was then analyzed with respect to these variables and tested for the hypothesized country and language level differences (research question 2). To be able to generalize the findings from the case study, the individual variables were prioritized by determining which ones showed the most significant country and language level differences, therefore indicating critical features for multilingual information systems design (research question 3). Finally, country profiles were developed as a tool to visualize different characteristics in comparison.

The three main research questions should be understood as stages of research, where one research question draws on the results of the previous one:

RQ1: Which variables in log files can be leveraged to study the user's country and language context?

RQ2: Does usage data indicate country or language specific interaction patterns?

> $\mathbf{H_0}$: Sessions from different countries and language backgrounds show the same interactions.
>
> $\mathbf{H_1}$: Country and language level differences exist between sessions.

RQ3: Which variables gathered by log files uncover significant country and language specific differences in user interactions?

While the first research question asks for potential candidate variables for studying country and language level differences, the third research question aims at validating these candidate variables for their significance and future use in other studies.

This staged research approach (figure 1.1):

- defines variables that signal country and language level differences in digital libraries,
- develops a logging format to trace these variables,
- determines an appropriate analysis method for characterizing differences,
- arrives at generalizable statements about significant factors, and
- visually represents the variables for country and language level comparisons in country profiles,
- which allow recommendations and directions for multilingual information access strategies to be provided.

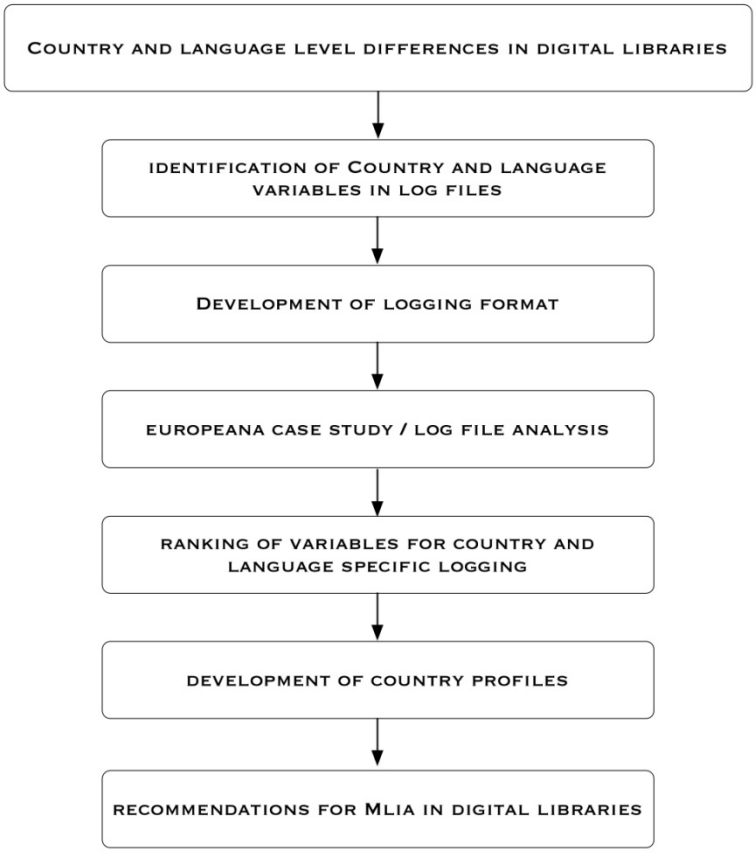

Figure 1.1 Structure and research stages of the dissertation

The main contributions of the study are summarized:

- Variables in log files are identified considering the country and language context in digital libraries from the user, system and content perspective.

- Application related information such as occurrence of facets as response to a search is not included in standard http logs. As this information can provide insights into user interactions and pathways through a system, a customized logging format was developed, the Europeana Language Logger, delivering extended information about the user and application under investigation.

- The dissertation presents the results of a deep log analysis of Europeana sessions as a case study. The Europeana portal provides a single access point to digital objects from Europe's cultural institutions such as libraries, audio-visual archives, museums and archives. Therefore, the multilingual digital library is an ideal use case for cross-country and cross-lingual studies. The thesis investigates 1,071,872 sessions from 21 countries. In total, 20 variables were considered with regard to country and language level differences.

- The dissertation evaluates the impact of each variable proposing a set of high and low impact variables for the investigation of the user's country and language context and differences. Out of 20 variables, 11 are classified as high impact indicators.

- Based on the identified country characteristics, a single profile is designed and graphically presented for each country. Exemplary comparisons are drawn between two individual countries, an individual country to an averaged country profile, content-rich and content-poor countries and English and non-English countries (averages over individual profiles).

- Based on the findings from this study, recommendations for multilingual information access to digital libraries are developed.

The quantitative methodology and analysis provided in this dissertation can serve as a basis for future studies of country and language level differences.

1.4 ORGANIZATION OF THE DISSERTATION

The content of this dissertation is organized as follows: Chapter 2 reviews previous literature and research in the domain of multilingual digital libraries and users as the object of study. The chapter provides an overview of the different levels of multilingual access to digital libraries with selected digital library projects as well as related user studies. The main outcomes and recommendations from previous studies focusing on aspects of multilingual information access are summarized and research gaps discussed.

Chapter 3 describes Europeana as a case study, which is used as the object of study in this dissertation. The system description is complemented by a language specific analysis of the available content, potential user groups and tasks within this multilingual digital library.

In chapter 4, the research background for the method of study is introduced. The focus lies on log file analysis as an unobtrusive method to measure user interactions including a discussion on strengths and limitations of this methodology. Units of analysis such as the query and session level as well as general metrics are discussed with respect to digital library applications.

Chapter 5 presents the specific logging approach developed and used for this study, including a description of the applied variables, selected countries, languages and statistical tests for this study. Based on the required user context information, direct and indirect indicators for country and language level differences provided by log file data are identified serving as a basis for the logging approach. For the purpose of this study, the Europeana Language – Logger (ELL) and its characteristics are explained. A corresponding log analyzer gathers specific statistics to identify country and language specific interaction patterns.

Chapter 6 presents the results from the log file analysis of 1,071,872 sessions from 21 countries. In total, 20 variables are investigated with regard to country and language level differences within the three digital library components: multilingual interface, multilingual search and browsing as well as multilingual result representation. For all variables and in particular for interface and result related interactions, significant differences between the countries are shown. The most significant differences are observed for the usage and preference of native language interfaces as well as for the refinement and selection of native language content. Based on the available content within Europeana, a differentiation of content-rich and content-poor countries is proposed. Finally, the applied logging approach is validated, proposing a set of high and low impact variables for the investigation of the user's country and language context and differences. From the 20 variables, 11 are classified as high impact indicators. The strongest

distinctive features are the usage of the Europeana interface language (change) as well as the usage of (native) language facets and content.

Based on the identified country characteristics, summary country profiles are designed and graphically presented in chapter 7. A comparison is drawn between two individual countries, an individual country to an average country profile, content-rich and content-poor countries and English and non-English countries (medians over individual profiles). Based on the country data and visualization options, several other comparisons and presentations are possible.

Chapter 8 summarizes the main outcomes and highlights additional findings and recommendations for MLIA. The dissertation concludes with an outlook on future and complementary work in the field of user studies in multilingual digital libraries. The focus lies on purposeful correlations, the impact of the interface language change and native content on user interactions.

2. MULTILINGUAL DIGITAL LIBRARIES

This chapter provides a review of previous work in the field of multilingual information access in digital libraries. The three main components or functionalities of multilingual digital libraries - (1) multilingual interface, (2) multilingual search and browsing and (3) multilingual result representation - are explained. Selected variables belonging to one of these components indicating country and language specific interactions are identified in chapter 5 and analyzed in chapter 6. The theoretical introduction is followed by a description of selected digital library projects and implementations demonstrating at least one aspect of multilingual information access. Related studies focusing on users within multilingual digital libraries are reviewed with regard to their main outcomes. The chapter concludes with a discussion of findings from previous user studies in multilingual digital libraries as well as the determination of current research gaps, some of which are addressed in this dissertation.

2.1 COMPONENTS OF MULTILINGUAL DIGITAL LIBRARIES

Multilingual access to information systems is a complex research area including system, user and business issues (Peters and Picchi, 1997; Peters et al., 2012). The following section provides an overview of multilingual functionalities within digital libraries. Multilinguality in search-based digital libraries has at least three component or functionality layers. The basic layer is the (1) multilingual interface that serves as a surface for the two main components (2) multilingual search and browsing functionalities and (3) multilingual result representation functionalities.

Multilingual User Interface (MUI) / Localization and Internationalization of Interfaces. The design and usability of interfaces has been discussed by several researchers, providing guidelines and best practices (Resnik and Vaughan, 2006; Hearst, 2009; Wilson, 2011). The design of multilingual user interfaces poses additional challenges in supporting international users and cross-lingual search tasks. The implementation of multilingual user interfaces is at least a two step process. At first, it needs to be ensured that the source code is flexible with regard to linguistic or culture specific requirements (internationalization). Secondly, the actual customization for each supported language or country needs to be implemented (localization). In other words, internationalization is the basis for localization. Nevertheless, the two concepts are often used interchangeably. The World Wide Web Consortium (W3C)[6] has provided general definitions for both terms in the context of web usage (Ishida and Miller, 2006).

[6] http://www.w3.org/

Internationalization strategies ensure that software can be easily adapted to different countries or languages. They should be an elementary part of the system design process, guiding the development of source code that enables localization and international implementation. Language-inherent challenges are, for example, different writing systems. While the majority of European languages follows the Roman script, the representation of other languages like Russian or Chinese needs additional adjustments (Large and Moukdad, 2000).

The localization of a system contains customizations of date formats, symbols, icons and other culture specific elements. The simplest form of localization, i.e. the adoption to a specific language including the translation of static pages, links etc., is called “language skinning”. In addition, culture specific issues need to be considered. While some concepts are language and culture independent, others might be misunderstood, like the display of different date formats.

The most common and elementary level of multilinguality in digital libraries is the adaptation of the interface language (language skinning). Currently, two options for interface language changes are predominant. Active interface language change options include implementations where the users select their preferred language via drop-down menus or pictograms such as flags. In contrast, passive interface language change options automatically determine the user’s language based on background information such as country information from the IP address or browser / agent language settings. Both alternatives pose advantages and disadvantages. Reducing the user effort with automatic geo-location fails whenever people are located in foreign countries. User-triggered interface language changes require an additional interaction, however. Observations in log files have shown that the system language is sometimes equated with the interface language by users causing confusion during the user-system interaction (Stiller and Gäde, to be published).

Multilingual Search and Browsing. Providing effective access to heterogeneous content in different languages is one of the main challenges of digital libraries. Overcoming the language gap between the user and content requires additional support functionalities. The most essential component of an actual multilingual information system is the cross-language search and browsing support. Classical search includes the formulation of a query and usually follows a more structured scenario while browsing activities vary from structured to explorative. Cross-language tasks include searching in a foreign monolingual collection as well as browsing multilingual content.

Depending on user and / or system requirements, multilingual search functionalities can be implemented in different ways:

1. Query translation: the original query is translated into the additional languages the document collection contains.
2. Document translation: the collection's documents are translated into the query language.
3. Pivot translation: queries and documents are translated into one language, the pivot language (Oard and Diekema, 1998; Oard, 1997; Jones et al., 2007).

While every approach comes with advantages and disadvantages, query translation has been the most commonly used solution due to its flexibility towards language changes (Agosti et al., 2009b). The query translation process includes several stages such as query formulation, reformulation, disambiguation, language detection and translation.

The challenge in processing queries in different languages includes the disambiguation of terms as well as named entity recognition. For example, the polysemous German query "Bank" has two different meanings and can be translated into "bench" (seating) or "bank" (financial). Depending on the underlying information need, different translation candidates should be displayed. For named entities, language independent names like "Albert Einstein" need to be recognized as such and excluded from the translation. Language dependent names need to be adapted and translated to the specific language version (e.g. "Spain" (EN) – "Spanien" (DE)). Other named entities occur in completely different versions only sharing the semantic correlation (e.g. Mona Lisa (DE) - La Jaconde (FR)).

When moving to a multilingual environment, the interaction can become very complex. Figure 2.1 displays a query translation prototype developed for Europeana, the digital library used as a use case in this dissertation. The example shows a search for "*storia del rinascimento*" with a

user determining the query language (Italian). For both query terms, translation candidates in German, French and English are displayed. While the translation of the term "rinascimento" is relatively clear, the term *"storia"* produces different translation candidates. The selection of an appropriate term is crucial for following the retrieval process.

Figure 2.1 Query translation prototype developed for Europeana

Depending on the implementation, this process is either hidden or user-assisted. From the user point of view, multilingual interactive information retrieval contains several interaction steps: the user determines the source language and / or the target language, the query translation process includes the examination of translation candidates and the possibility to add alternative translations, and finally, the displayed results can be sorted by language and translated into the user's preferred language(s). It is an open research issue how and to what extent systems can support the query translation process, especially with ambiguous terms (Petrelli et al, 2008; Oard et al., 2004).

In addition to traditional search support, digital libraries aim at providing alternative access through classification or category systems and ontologies. Multilingual browsing is an essential feature for users who do not feel comfortable searching in foreign languages or want to discover unknown content and context. Structured data can be explored via browsing paths, linking related objects or topics, and tag clouds or time lines. Particular challenges for the translation of classifications or other category systems are culturally diverse concepts and representations (Soergel, 1997).

Primarily designed for children, parents and teachers, the International Children's Digital Library has developed alternative search and browsing options to meet their user requirements and to circumvent some of the multilingual challenges (figure 2.2). For example, users can start their search via a color facet if they only remember the booklet color from their childhood or via a length facet. Similarly, users can browse through topical collections or select content according to their preferred language or age ranges finding the appropriate books for individual reading skills.

Figure 2.2 Search and browsing interface provided by the International Children's Digital Library

Result Representation. Multilingual result representation is concerned with the organization of results in different languages. For example, retrieved objects can be displayed either in a common ranked list or separated by their language. Different languages require different displaying options. While most languages are left-right oriented, other languages like Arabic need to be presented from right to the left.

Individual preferences for the presentation of multilingual results could be determined through personalization settings. Some systems request users to specify their location, language and result preferences when they create their user profile. This information can be used to customize the interface as well as system settings.

Advanced search fields or facets (filters) are options to refine results by language or country of origin. Through the advanced search interface, users can select their preferred result language(s) and include and display only those for the search results. Another option would be to refine the complete result with all available languages set via facets.

Apart from the representation of multilingual results, users need further support to decide which results are relevant to their information need. Results can be translated at the metadata or the

digital objects level. For language independent objects like images and sound files, only bibliographic data is available for translation. Information provided by object "snippets" is often sufficient to examine a specific result. Depending on the user's language skills, different translation levels may be required. Usually, machine translation (MT) solutions of the metadata are preferred to expensive full text translation. User studies have shown that partial or imperfect translations of textual objects are still useful to examine the meaning of a document (Marlow et al., 2007).

2.2 MULTILINGUAL DIGITAL LIBRARY PROJECTS

Within the cultural heritage domain, a few applications that integrate multilingual design aspects have been developed. Due to the cultural and language diversity of the European Union, especially European applications apply multilingual access strategies.

Projects either focus on digitization, the collection and development of language resources and translation services, or on providing prototype systems. Especially long-term projects like The European Library (TEL)[7] and Europeana[8] as well as projects associated with them have been the focus of research. The selected projects or systems are discussed with regard to multilingual information access component implementations, presented in chronological order starting with the early implementations. Related user studies are discussed in the following sections with a summary of previous findings and recommendations in section 2.4.

2001 – 2003. A user centered design approach was applied for the Clarity[9] search engine providing two interface versions for query translation (Petrelli et al., 2004). Clarity focused on usability aspects related to multilingual interface and search or browsing functionalities. The supervised mode presented a highly interactive solution where the user could control the query translation process by judging and correcting wrong translations. In contrast, the delegated mode represented a fully automatic system setting without any user assistance (Petrelli et al., 2008).

*2002 - **. The International Children's Digital Library (ICDL)[10] was launched in 2002 with the aim to offer children's literature from all over the world providing books in several languages. Primarily designed for children, parents and teachers, the library has developed alternative search options and facets to meet their user requirements (Druin, 2001; Hutchinson et al., 2005).

[7] http://www.theeuropeanlibrary.org/tel4/
[8] http://www.europeana.eu
[9] http://web.archive.org/web/20081226001907/; http://www.dcs.shef.ac.uk/nlp/clarity/index.html
[10] http://en.childrenslibrary.org/

In 2013, the ICDL contains 4643 books in 61 languages and supports an interface in 19 languages. Summaries for each book can be displayed in the user's preferred language through a drop-down menu. The library maintains full texts and attempts to translate them with the help of volunteers. For simple and advanced searches, the target language can be selected beforehand.

2004 - *. Under the acronym Minerva[11] (Ministerial Network for Valorizing Activities in Digitization), a network was built to implement an infrastructure for digitization activities and best practices. The Minerva activities mainly focused on the provision of guidelines for the improvement of user interaction and satisfaction for cultural heritage websites (MINERVAplus, 2006). Within the project, a survey was conducted to identify and evaluate European multilingual cultural websites and their usage of multilingual thesauri. Most websites provide multilingual interfaces but only a few heterogeneous multilingual thesauri exist (Caffo et al., 2008).

*2005 - *.* Since 2005, The European Library (TEL)[12] has aggregated content of 48 European national libraries and leading European research libraries (Cousins, 2006). Currently, TEL supports 36 interface languages, allowing access to more than ten thousand digitized objects and almost 107 million bibliographic records. The portal offers a simple and advanced search interface as well as several categories to discover content such as Discipline, Content Language and Date of Publication. Users can either search the complete index or choose a provider and language beforehand. Within result sets, a set of facets can be used to refine lists. Depending on their accessibility, resources can be viewed in their original context or exported to reference management tools like Mendeley[13]. The development of TEL including the enhancement of multilingual access to the portal through a multilingual interface and language filtering of results was supported in multilevel implementations (Mane, 2009; Clavel-Merrin et al., 2006, 2008; Braschler and Ferro, 2007).

TEL also links to resources available at the Europeana portal. The Europeana portal offers access to digital objects from Europe's cultural institutions such as libraries, audio-visual archives, museums and archives. Currently, Europeana offers a multilingual user interface, country and language facets as well as document (metadata) translation via an external translation service. A more detailed description of Europeana can be found in chapter 3.

2006 – 2009. The MultiMatch project[14] developed a multilingual and multimedia search portal for unified access to cultural heritage material. The prototype[15] offers cross-language search and

[11] http://www.minervaeurope.org/
[12] http://www.theeuropeanlibrary.org/tel4/
[13] http://www.mendeley.com/
[14] www.multimatch.eu

browsing functionalities like metadata-based retrieval (Amato et al., 2007). The enhanced access strategies include a query translation module for the provided languages (Marlow et al., 2008a, 2008b).

2007 – 2009. The CACAO project (Cross-Language Access to Catalogues and Online Libraries)[16] was launched in 2007 in order to build an infrastructure for cross-language access to digital library content (Levergood et al., 2008). Focusing on query translation, necessary components and language resources for effective translation modules were identified and collected (Bernardi et al., 2009).

*2007 - **. In 2007, the Michael Culture Association[17] was founded to support and maintain the MICHAEL search portal. At the moment, the MICHAEL Multilingual Inventory of Cultural Heritage in Europe portal[18] offers a multilingual simple and advanced user interface in 17 languages (Fresa, 2005). The advanced search offers filtering options for every metadata field. Alternatively, users can browse by content, institution type or location or services such as audience, subject, coverage or period. Results can be translated by an external translation service and easily exported in several formats.

2.3 STUDIES DEALING WITH MLIA IN DIGITAL LIBRARIES

In line with their efforts on establishing effective access to their content, several digital library projects have conducted studies on user needs and requirements, but few have paid specific attention to multilingual information access (MLIA) issues. The studies vary in terms of research methods, including observations, surveys, interviews, task-based experiments or log file analysis. A comparison or generalization of findings is difficult because of different systems requirements, varying number of participants, the amount of usage data collected and several other factors. Although a clear separation of outcomes regarding the different components of multilinguality cannot always be made, it was attempted to distinguish the most important results related to the three DL levels: multilingual interface, search and browsing as well as result representation. A brief excursus dealing with studies focusing on the cultural and linguistic background of users completes the provided overview. A summary of previous findings and recommendations can be found in table 2.1.

[15] http://multimatch01.isti.cnr.it/overview/
[16] http://www.cacaoproject.eu
[17] http://www.michael-culture.eu/
[18] http://www.michael-culture.org/

The majority of studies were conducted in the context of a particular digital library or project. Case studies vary from domain specific repositories to large-scale digital libraries. Research questions either address a single aspect of MLIA or consider the complete application. Most studies included test subjects from an academic setting. Only a few researchers involved end users in their studies. In the broadest sense, each of the studies addressed at least one of the following questions:

- Who are the users of multilingual content or systems?
- How are users interacting with multilingual systems and for what reasons?
- Which features are needed to support multilingual access to information systems?

Subject to the available data and resources, researchers either applied qualitative or quantitative approaches. While several studies focused on a particular research method, others combined methods to complement their findings.

The most popular method was the online questionnaire or survey (Marlow et al., 2007; Bernardi et al., 2006, 2008; Clough and Sanderson 2006; Petrelli et al., 2002; IRN Research 2009, 2011; Wu et al 2012; Minelli et al., 2006). Studies making use of qualitative methods such as (expert) interviews, focus group discussions, think aloud tests and questionnaires have the advantage that user background information and preferences can be gathered. A few projects also organized workshops inviting researchers and stakeholders in order to determine user requirements or multilingual access strategies (Agosti et al., 2009b; Gonzalo et al., 2008; Minelli et al., 2006)

Quantitative methods such as log file analysis were mainly conducted in order to identify usage patterns to inform interface design as well as content development or enrichment. Especially for studies exploiting log files, no standard procedures and metrics could be determined. Several digital library projects such as Europeana (Clark et al., 2011), The European Library (Mandl et al., 2010, 2010a) or the CACAO project (Trojahn and Siciliano, 2009) provide and analyze web server logs as well as search queries with the intention to improve the portal functionalities according to the results derived from the analysis of user behavior.

To overcome the shortcomings of either qualitative or quantitative methods, some researchers combined data from complementary sources like performance observations, eye tracking records and interviews or questionnaires (Aula and Keller 2009; Agosti, 2010; Dobreva et al. 2010, 2010a; Marlow et al., 2007; Srinivasarao 2008; Bernardi et al., 2008; Angelaki, 2007; Bilal and Bachir, 2007, 2007a; Clough and Eleta, 2010).

2.3.1 THE USER'S CULTURAL AND LINGUISTIC BACKGROUND

Language is an essential part of the user's cultural identity. Although this study focuses on the user's origin and assumed native language, a brief overview of studies dealing with cultural and linguistic issues is presented here.

Cross-cultural human computer interaction (HCI) research researches the relation between culture and information systems. It mainly applies the cultural models and dimensions developed by Hofstede (1983, 2010) or Hall and Hall (1990). Based on this work, several guidelines and frameworks for international interfaces and system design were prepared (Zahedi et al., 2001; Hsieh et al., 2009; Marcus and Baumgartner, 2004; Ford and Kotze, 2005, 2005a; Ford and Gelderblom, 2003; Jones and Alony, 2007) but only a few evaluated their proposed models through usability studies (Hsieh et al., 2009; Markus and Alexander, 2007). The majority of cross-cultural HCI studies focuses on the examination of interface issues investigating the influence of cultural background, arguing that cross-cultural research needs to focus on interface representations rather than on "simple" language skinning (Bourges-Waldegg and Scrivener, 1998).

Some studies broadened their scope to the examination of user behavior and expectations, but the results are inconsistent (Rau et al., 2004). An experiment on Hofstede's cultural dimensions and their impact on human computer interaction by Ford and Gelderblom (2003) could not find a significant correlation between cultural dimensions and user performance.

In a number of papers, Anette Kralisch studied the impact of language, culture and domain knowledge on user behavior and navigation strategies (Kralisch 2005; Berendt and Kralisch 2009; Kralisch and Mandl, 2006). Kralisch assumes that differences in search behavior are based on the level of cognitive effort the user needs to access a website. Consequently, language skills and domain knowledge should influence search strategies and preferences for different search options such as search engines, alphabetical search and hyperlink navigation.

Few studies have examined cultural based usability issues related to digital libraries (Caidi and Komlodi, 2003; Komlodi et al., 2004). Even less studies target the influence of the user's language on information access. Studies are often limited to the comparison between native and non-native participants or focus on one cultural or regional group (Bilal and Bachir, 2007, 2007a). Usually, English is used as a baseline and compared to another language.

2.3.2 MULTILINGUAL USER INTERFACES

Numerous experts point out the importance and consistency of localization and internationalization of user interfaces (Directorate-General Information Society and Media, 2011). Although users express their preference for international interfaces (Agosti et al., 2009b), analysis of usage data has shown that the majority of users stay with the default interface language English (Angelaki, 2007; Agosti et al., 2007). For those sessions that contain an interface language change, it is most likely that users switch to their native language (Agosti et al., 2007).

Another important research question is the influence of the interface language on search strategies, satisfaction and success. Keegan and Cunnigham (2005) investigated the impact of interface language change between Māori and English on usage patterns. Using the example of the New Zealand Digital Library (NZDL)[19] and choosing a predominant Maori collection, they analyzed usage logs of four weeks, switching the interface language settings between English and Maori each week. Users are more likely to conduct a search when the default interface language is set to English. Users seemed to feel comfortable with the English interface when it was the default version and showed a strong preference for the English interface during the "Maori weeks". During the "Maori weeks", only 26% of all sessions were conducted with the native interface but 74% switched the interface language back to English. Although English as interface language was preferred, Keegan and Cunnigham observed that Maori interface sessions retrieved more results and showed a higher request number for full result views. Additionally, Maori sessions showed a stronger preference for browsing or navigational access while English sessions are more search dominated (Keegan and Cunnigham, 2005).

2.3.3 MULTILINGUAL SEARCH AND BROWSING

Language as an essential part of the user's context and information gathering is still an underrepresented topic in research studies within the digital library domain. In the context of MultiMatch, an extensive user requirements analysis was conducted to identify user groups, their individual needs and potential scenarios within the cultural heritage domain (Minelli and Naldi, 2006). The Minerva handbook on cultural web user interaction gives an overview of needs and expectations of cultural heritage user groups, describing possible multilingual access strategies (Caffo et al., 2008). In addition, log files were gathered to observe user behavior and preferences (Minelli et al., 2007).

[19] http://www.nzdl.org/

While it is not surprising that users feel more comfortable searching in their mother tongue, they frequently search in other languages as well (Aula and Kellar, 2009; Clough and Sanderson, 2006). In 2011, language preferences of EU internet users were surveyed with the result that more than half of all participants (55%) used at least one additional language, mainly English, besides their native language when accessing content in the web (Directorate-General Information Society and Media, 2011). In particular, browsing related activities were often performed in other languages (81%).

Especially if the initial (native) query was unsuccessful or broader results are expected, users tend to repeat queries in another language (usually English) (Srinivasarao et al., 2008; Aula and Kellar, 2009; Marlow et al., 2008). Other studies found a dominance of English queries - only 24% of all queries matched the native language(s) of the country they were submitted from (Leveling et al., 2010). Through a study of search logs, the CACAO project found that in a multilingual library environment, about 20% of the queries were written in three languages, namely Italian, German and English pointing to the multilingual capabilities of their users (Trojahn and Siciliano, 2009). A search study with users from different language backgrounds observed that 20% of term changes involved language changes (Ghorab et al., 2010), while others found that users rarely switch the query language during a search session (Oakes and Xu, 2009).

The needs and expectations of academic users from 19 different countries within multilingual digital libraries were studied by Wu et al. (2012). Most respondents (84%) expressed the need for domain specific term translation functions as well as cross-language search and browsing options. The analysis also showed that users from different countries and language backgrounds express different needs or expectations for multilingual access features in digital libraries. Especially non-English users had experiences with translation tools and strong preferences for multilingual information access (Wu et al., 2012).

Search assistance and interactive information retrieval functions are even more important when users have to deal with content in multiple languages. Performing different tasks with an interactive multilingual prototype, participants expressed the need to choose the language they want to search in, depending on the individual skills and the task (Petrelli et al., 2002). Based on their findings, the authors suggest that user-assisted query translation should be offered as an advanced search option if the initial query translation fails or does not satisfy the user's information need. A survey focusing on multilingual access to Europeana found that the majority of users (80%) were willing to control the query translation process (Agosti et al., 2009b).

In 2008 and 2009, the interactive CLEF track iCLEF focused on problems of multilingual search assistance. A multilingual search interface in 6 European languages (EN, ES, IT, DE, NL, FR) provided monolingual and multilingual search options with a high amount of interaction possibilities such as query refinement and the inclusion of translation suggestions by the user (Peinado et al., 2008). In both years, 5,101 search sessions were logged from 40 different countries to investigate user behavior in a multilingual environment (Gonzalo et al., 2008). In most cases, users showed a preference for their native language, preferring the monolingual interface and only switching to the multilingual one if they realized that cross-language search was necessary to find a given image. Those who often reformulated their queries instead of looking at many result pages found most images (Srinivasarao et al., 2008). Tanase and Kapetanios (2008) focused on the relationship of language skills and the use of interaction features such as translation modification and personal dictionaries. Users with lower language skills added query translations more often than users who spoke several languages. Peinado et al. (2008) classified search sessions according to active, passive or lacking language skills of the user and studied the behavior of users during a session. Sessions belonging to the active and passive group did not vary in success rates but those users with passive language skills made higher efforts. Participants from both groups found it hard to choose the appropriate translations but got familiar with the system features and needed less effort during the session.

Comparing the different results for monolingual or multilingual searches, no significant differences could be found; both variants as well as the mixed approach did not influence the success rates of users (Di Nunzio, 2008). An in-depth study showed that 8 out of 10 users switched between the monolingual and the multilingual interface. Those that only used monolingual settings justified their decision with a lack of language knowledge. Only 4 out of 10 users realized the important role of language for retrieving known items. The findings from questionnaires and interviews demonstrate that most users do not consider the relationship between the query language and the language of the object they are searching for (Vassilakaki et al., 2009). On the other hand, it was shown that knowledge of the target language improves the users' success rate by 12%. Another interesting result was that users with active and passive language skills do not significantly vary in success rate but in interaction (Peinado et al., 2008).

The LogCLEF track was launched in 2009 with the aim to study user behavior in multilingual search systems through the analysis of activities and search queries. In 2009 and 2010, log files from different providers were evaluated in order to understand search behavior in multilingual contexts and to improve search systems (Mandl et al., 2010, 2010a). Using TEL log files, a special focus was put on the query language identification for search result improvement. The high amount of queries for named entities in the cultural heritage domain can often not be

assigned to one language and poses a challenge when using machine translation solutions (Stiller et al., 2010). The inclusion of session information and additional language indicators such as the origin of users or the selected interface language could not deliver satisfactory results in determining the language of a query (Gäde et al., 2011).

The influence and importance of the user's language background on search behavior was also studied related to other factors like gender, age or domain knowledge and academic background. Studying search behavior of native and non-native male and female students, Zoe and DiMartino (2000) found no significant difference between gender groups but an influence of language background on search strategies, satisfaction and success. They observed a strong correlation between native language sessions and satisfaction.

In the context of the International Children's Digital Library, the understanding of interface representations as well as information seeking behavior of Arabic speaking children was studied (Bilal and Bachir, 2007, 2007a). They argue that design representations need to be intuitive irrespective of the user's cultural background and the provided interface language(s) (Bilal and Bachir, 2007a). All participants made use of the language facet "Arabic" to search for native content rather than using free text, advanced or location search (Bilal and Bachir, 2007a).

Domain knowledge of a specific field has been identified as an important factor influencing search behavior within multilingual environments (Clough and Eleta, 2011). Depending on the need and integration of foreign information, different disciplines show different usage patterns and language preferences.

Research comparing native and foreign web users has shown differences in search behavior. While foreign searchers spend more time and frequently reformulate queries, no significant difference was observed for search engine result pages (SERP) viewed, websites clicked and navigation on retrieved websites (Chu et al., 2012). In contrast, Józsa et al. (2012) observed differences in information seeking strategies between native and non-native search tasks. Users searching in their native language tend to visit more websites than users scanning websites with non-native content. The inspection and judgment of foreign content is one of the biggest challenges for multilingual information access.

Using click stream data from a search engine, Gandal and Shapiro (2001) investigated differences between native French and English speakers. For their study, a region with two official languages was chosen (Canada / Quebec). For every result view, the language as well as the content of the retrieved website was determined. Additionally, demographics and language skills of the participants were included. They found no significant difference concerning the

frequency of internet usage between French and English speakers. Concerning website usage, they observed preferences for government sides from French speakers whereas search engines are more popular within the English speaking group. Younger French users (under 15 ages) with a good knowledge of English tend to spend much more time on the internet than English speaking children.

2.3.4 MULTILINGUAL RESULT REPRESENTATION

Studies dealing with multilingual result representation and translation mainly consider preferences for and interaction with result sets in different languages. In general, users feel comfortable accessing a portal and scanning results in their native language or in English but struggle when dealing with content in unknown languages (Dobreva et al., 2010).

While EU internet users expressed their wish for native content (Directorate-General Information Society and Media, 2011), web search users rarely indicate preferences for result languages, but rather search for all available websites. Studying search patterns from European AlltheWeb.com users, most language specifications for results were observed for French, Spanish and German (Jansen and Spink, 2005). In the context of TEL and Europeana, it was found that users tend to select collections that correlate with their native country (Agosti et al. 2009, 2009a; Clark et al., 2011). Agosti et al. (2009a) observed similar preferences for users from one country group. For example, Spanish, Italian, French and Canadian users all preferred French collections, while users in Germany, Poland, Hungary and Croatia preferred German content. A Europeana online survey determined that the most popular result refinements are the language and country facets (IRN Research, 2009).

Dealing with non-native content, users are more likely to visit collections if they are translated into their preferred language. Additionally, they are even willing to accept a text that they could understand but which was not perfectly translated (Wu et al., 2011; Clough and Sanderson, 2006; Minelli et al., 2006; Minelli et al., 2007; Gonzalo et al., 2008).

Leveling et al. (2010) investigated the influence of the user's origin as well as the interface and query language on collection selection. Based on this information, alternative re-rankings of collections were produced listing native language and country collections higher, showing an overall improvement in precision.

Only a few studies investigated country based paging behavior. A LogCLEF study observed extended result list paging behavior for users from Britain, Italian, Poland and Spain while other language groups tended to reformulate queries to find relevant content (Lamm et al., 2010).

Best practices recommend that systems should at least provide two options for result representation: either merged or separated by language (Gonzalo et al., 2008).

2.4 PREVIOUS FINDINGS AND RESEARCH GAPS

Table 2.1 summarizes the main outcomes and recommendations for the different levels of multilinguality in digital libraries which were researched in the studies described above.

Component	Outcomes	Recommendation	Reference
Multilingual Interface			
	Users feel more comfortable when they can access websites in their native or preferred language and predominantly switch the interface language to their native language.	Provide multilingual interface in all supported languages.	Agosti et al., 2007; Agosti et al., 2009; Dobreva et al., 2010
	Users rarely switch the interface language but commonly accept the default language English.	Provide English interface as default or automatic interface language change by default.	Angelaki, 2007; Agosti et al., 2007; Clark et al., 2011; Oakes et al., 2009; Keegan and Cunnigham, 2005
Multilingual Search and Browsing			
	User struggle when dealing with multilingual query translation.	Offer user-assisted query translation only if automatic translation fails. Translation candidates should be limited to avoid user effort.	Petrelli et al., 2002; Gonzalo et al., 2008
	Users want to control the query translation process.	Provide advanced search functionality for user-assisted query translation.	Agosti et al., 2009; Gonzalo et al., 2008
	Users want to search in their native language but	Provide cross-language retrieval.	Srinivasarao et al., 2008; Aula and Kellar,

	repeat queries in another language (usually English) if the initial (native) query was unsuccessful or broader results are expected.		2009; Trojahn and Siciliano, 2009; Ghorab et al., 2010; Leveling et al., 2010; Marlow et al., 2008; Directorate-General Information Society and Media, 2011
	The user's language background influences information needs, search strategies, satisfaction and success.	Provide different search assistance options.	Lamm et al., 2010; Zoe and DiMartino, 2000; Bilal and Bachir, 2007b; Peinado et al., 2008; Keegan and Cunnigham, 2005; Wu et al., 2012
	Users do not understand the relation between query and object language.	Provide clear descriptions of the functionalities and limitations of multilingual search.	Peinado et al., 2008
Multilingual Result Representation			
	Users prefer different result representations depending on language skills and information need.	Provide at least two options for multilingual result display (merged or separated by language).	Gonzalo et al, 2008
	Users frequently refine results by language.	Provide language refinement via advanced interface and facets.	IRN Research, 2009; Bilal and Bachir, 2007b
	(Some) countries show high preferences for native content.	Consider higher ranking of native content.	Clark et al., 2011; Agosti et al., 2009a, 2009b; Leveling et al., 2010; Directorate-General Information Society and Media, 2011
	Metadata translation or summaries are sufficient.	Provide machine translation for	Oard et al., 2004; Gonzalo et al., 2008;

	Users should at least have the possibility to scan the content.	metadata, provide easy switching between original object and translation.	Minelli et al., 2006; Clough and Sanderson, 2006
Personalization			
	User interactions provide background information and preferences.	Leverage information from previous sessions, provide storage and re-use of language preferences and skills in the user profile.	Gonzalo et al., 2008; Saulnier and Viand, 2009; Agosti et al., 2007

Table 2.1 Outcomes and recommendations for multilingual information access

The previous findings show a serious interest in MLIA issues. It has been proven that users with different language skills have different needs and expectations accessing digital libraries. While the user's language and country background are considered important factors for digital library access, less has been done to identify differences in interactions based on these factors that can be leveraged to provide effective systems.

The majority of studies exploited qualitative usage data focusing on a small group of users and their requirements. Studying international users comes with the challenge of representative user groups requiting participants from all over the world requiring user studies in several languages. Assembling quantitative usage data as provided in log files allows observing trends from country groups in their natural environment. So far, international usage data from a digital library has not been investigated or interpreted with regard to country or language level differences.

2.5 SUMMARY

Multilingual digital libraries are complex systems with an interplay of functionalities belonging to one of the three components (1) multilingual interface, (2) multilingual search and browsing and (3) multilingual result representation. The widespread implementation of multilingual user interfaces contrasts with a few applications that provide multilingual search and browsing as well as multilingual result representation functionalities. Therefore, the majority of studies in this field have dealt with the usage and acceptance of interfaces. Some studies have focused on

single aspects such as language skills and multilingual search tasks observing a small sample of users from limited countries.

Recent studies and projects point out that the user context information and usage data should be leveraged in order to personalize search experience. Initiatives like the PATH (Personalised Access to Cultural Heritage Spaces)[20] project and the CULTURA (Cultivating Understanding and Research through Adaptivity) [21] project promote personalized system design focusing on individual user contexts. While researchers emphasize that language preferences indicated in user profiles should be adapted and remembered for future visits, only a few systems have implemented personalized search (Saulnier and Viand, 2009). The same is true for user-assisted translation functionalities.

This dissertation proposes an in-depth log file analysis method interpreting user interactions with a special focus on country level differences and native language preferences focusing on all three levels of multilinguality in digital libraries. As a case study, the Europeana portal serves as an instrument for the analysis of country and language level differences between international users in multilingual digital libraries. The results derived from this study will be reviewed in the context of previous recommendations for multilingual information access in chapter 8.

[20] http://www.paths-project.eu/
[21] http://www.cultura-strep.eu/

3. CASE STUDY EUROPEANA

The Europeana portal aggregates content from Europe's national libraries, archives, audio-visual archives and museums. Because of its multilingual content as well as its international audience, Europeana is especially suitable to study country and language level differences in digital libraries. This chapter introduces Europeana as a case study for analyzing country and language level differences. Europeana's strategic objectives, its users, content and multilingual functionalities are introduced in order to demonstrate its suitability as a multilingual digital library analysis object.

The description and screenshots demonstrate the portal version during the data collection period from August 2010 to May 2011. Since then, several interface and system changes have been implemented. Stiller et al. (2013) provide a detailed description and evaluation of recent implementations and developments within Europeana.

3.1 EUROPEANA'S MISSION AND OBJECTIVES

As the European reaction to Google's digitization statement, the Europeana prototype was launched in 2008 with the vision to allow boundless access to Europe's digital cultural heritage material (Purday, 2009). Founded by the European Commission, the European Digital Library Network (EDLnet) initiated the preparation of TEL and Europeana with a special focus on multilingual access functionalities. Many projects participate in the development and improvement of the portal's functionalities and content[22]. Within the technology oriented project EuropeanaConnect[23], multilingual access strategies and technologies were developed (Agosti et al., 2009b; Petras, 2011). The main development phase from the Europeana prototype to an operational system was managed by the Europeana v1.0 project[24] (2009 – 2011).

Europeana's strategic plan for 2011 to 2015 defines four main steps that can be seen as the main pillars for Europeana's future advancement (Europeana Foundation, 2011). In particular for the distribution of cultural heritage material, multilingual information access plays an important role.

Aggregate: Through the consistent aggregation of new content (providers), Europeana expands its coverage. Another goal is to improve and contextualize metadata descriptions of Europe's

[22] A list of projects is provided by the Europeana Foundation: http://pro.europeana.eu/web/guest/projects;jsessionid=4728D8F799FDAB9874EC027A960E2422
[23] http://www.europeanaconnect.eu/
[24] http://pro.europeana.eu/web/europeana-v1.0

digitized cultural heritage content with the provision of standards like the Europeana Data Model (EDM).

Facilitate: The Europeana project group combines the interests and needs from professionals, managers and developers providing a platform for continued knowledge sharing. Europeana promotes the development and usage of open source products, for example offering the Europeana API (Application Programming Interface) to the community. Additionally, the Europeana network frequently organizes events bringing together people from the cultural sector as well as researchers and stakeholders.

Distribute: Europeana's main goal is the boundless access to cultural heritage objects irrespective of the user's origin, location, language or device used. Through alternative access modes, the user experience will be improved.

Engage: Europeana wants its users to actively engage with Europe's cultural heritage. The goal is to establish a platform that involves users in the creation of virtual exhibitions and encourages them to add content and context through tagging services, storytelling or ratings. This will be mainly achieved through personalization options, enabling users to customize the portal according to their needs.

The portal has become an important authority for the maintenance and preparation of digitized cultural heritage objects. Currently, the Europeana Foundation[25] is responsible for the Europeana services, supporting the constant refinement and improvement of the system and the available content.

The following sections provide an overview of the Europeana portal from the system, content and user perspective, highlighting why this particular portal was chosen as object of study.

[25] http://www.pro.europeana.eu/about/foundation

3.2 SYSTEM

The Europeana system description is structured according to the multilingual digital library components identified in chapter 2.

Multilingual Interface. The Europeana user interface is provided in 29 languages. The localization function allows users to page through all static pages in the selected language. For Europeana, three types of interface language change options can be distinguished (Gäde and Stiller, 2011):

- Link change (via external referrer),
- User change (drop-down menu),
- Cookie change (automatic invisible change).

Users that do not directly type in the Europeana URL usually access the portal via external links. Search engines present a default link as well as a link requesting the local version of the website. In general, both versions are available at the first result page and users can consciously decide which link they prefer. Figure 3.1 shows a German Google result page for the query "europeana" presenting the default (English) version at first position followed by the German adaption (www.europeana.eu/portal/?lang=DE).

Figure 3.1 Google result page with default link and German language version

A first time visitor can make use of the drop-down menu to choose another interface language than the default English one (figure 3.2). All static pages are translated according to the selected language for the duration of this particular visit.

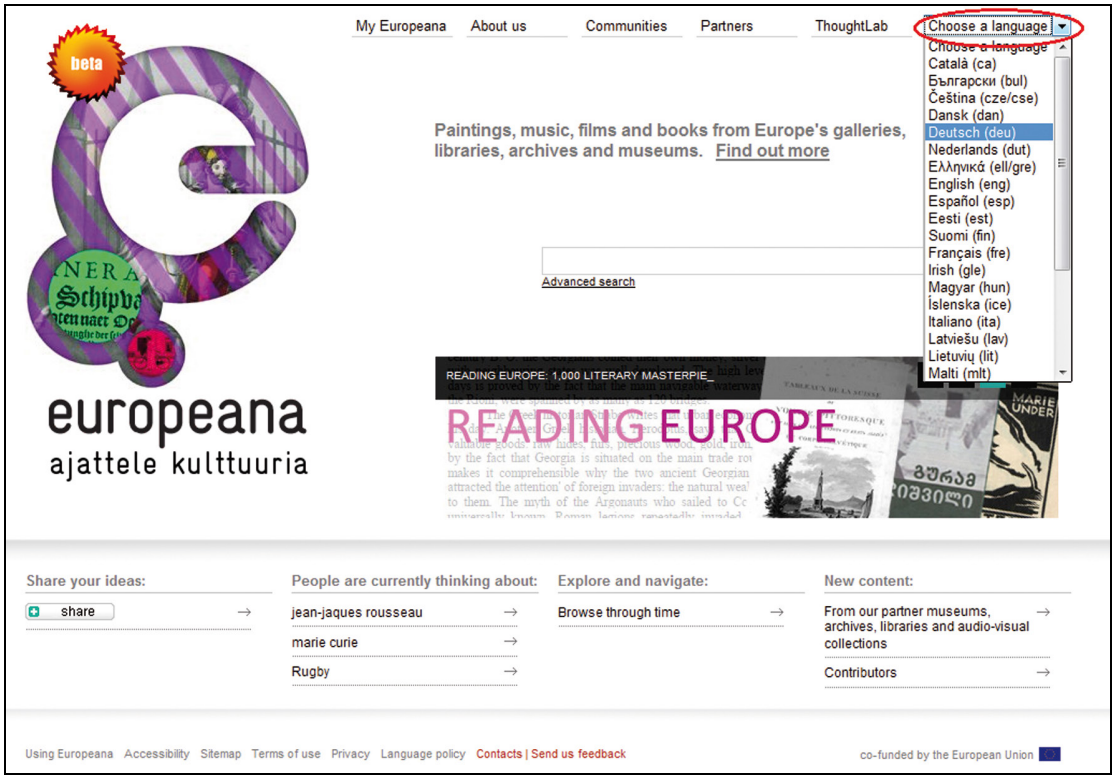

Figure 3.2 Multilingual interface languages – drop-down menu

Once a user has made a language choice via the drop-down menu, this information is stored in cookies. In this way, repeat visitors and their language preferences are identified and automatically changed.

It is also possible that different language change options are combined. A returning visitor will be automatically directed to their preferred language version. During the same session, this user might change the interface language again via the drop down menu (cookie and user change). Another interface language combination would be a user accessing the portal via a non-default link version again followed by a user change via the drop down menu (link and user change). In chapter 6, the frequency of interface language change (types) is presented.

Multilingual Searching and Browsing. Europeana only offers monolingual search functionalities. Users that want to receive results in different languages need to enter the same query in several languages. For example, the query "Eiffelturm" would only return results where the German search term "Eiffelturm" matches the metadata description. If a user is also interested in French objects, it is necessary to repeat the query in French ("Tour Eiffel"). The

advanced search functionality allowed users to specify their query through additional fields like author, title or language. Due to the poor usage of this feature, the advanced search options were integrated into the simple search interface allowing users to search all fields or select a specific one.

Up to date, the Europeana portal added features like the auto completion suggesting terms in different languages that might be related to the user query. Initial searches can be refined by additional searches if too many results were found. However, cross-lingual search is still an open issue.

Besides the search functionalities, visitors could use browsing features to navigate and explore the content. Virtual exhibitions as well as a timeline display are an alternative way to access content without the need to enter search terms. The most popular browsing feature provided during the data collection period was the "people are currently thinking about" (PACTA) presentation. Using previous search queries, manually determined query suggestions were presented to the user at the homepage. According to the interface language, the suggested query terms are displayed in the appropriate language (see section 6.2.3).

Multilingual Result Representation. Figure 3.3 shows a result page for the query "peter pan" which resulted in 159 retrieved objects. Result lists are displayed in four media type categories: text, image, video and sound. In this case, 31 texts, 111 images, 6 videos and 11 sound files were found.

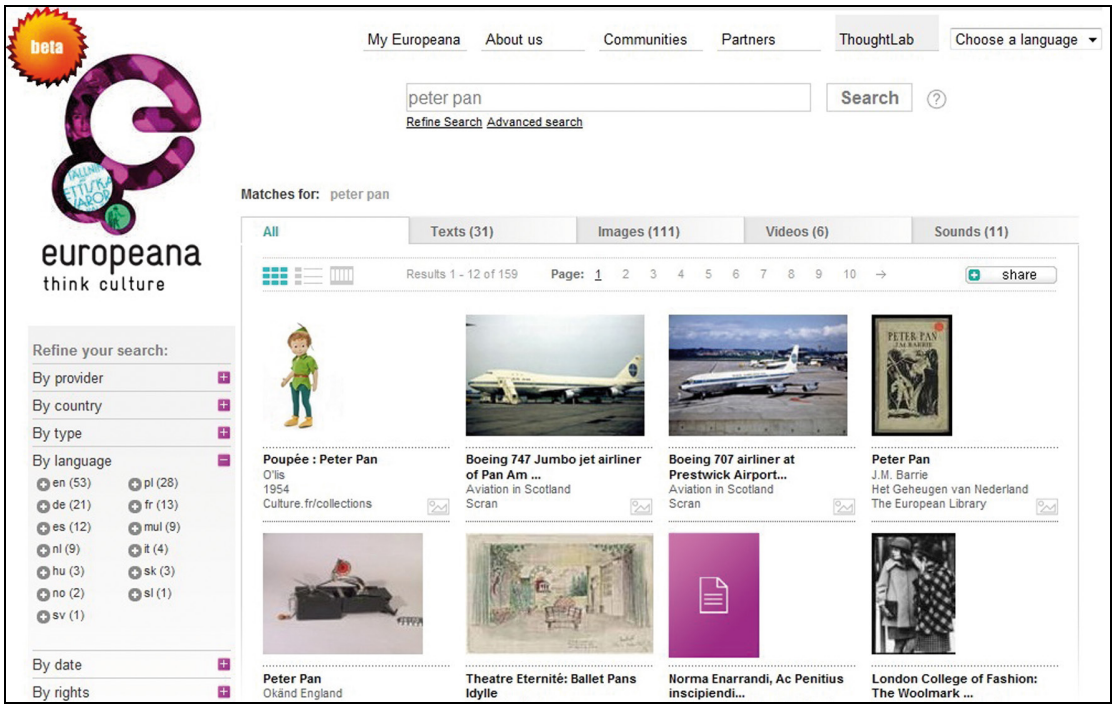

Figure 3.3 First result page for query "peter pan"

In addition to the traditional grid view, results can be displayed on a timeline arranging objects by date of origin. The result list can also be refined through facets. In total, five facets are available: provider, country, type, language, date and rights. Through the language and country facet, one or several languages and / or countries can be specifically selected to reduce the result set. The country facet presents the country of the content providers and does not necessarily provide information about the actual country of origin of the object or its language, since some collections contain content in several different languages. The same is true for object language. For Europeana, the object language is determined by the content provider and does not necessarily refer to the object's metadata language. In other words, the Mona Lisa appears in several languages depending on the hosting institution.

If a user clicks on an object, further details are displayed as well as a link that leads to the actual content provider and a more detailed view of the object or the object itself. During the data collection period, Europeana offered metadata translation via the external Google and Bing Translate services on this full object view page (figure 3.4).

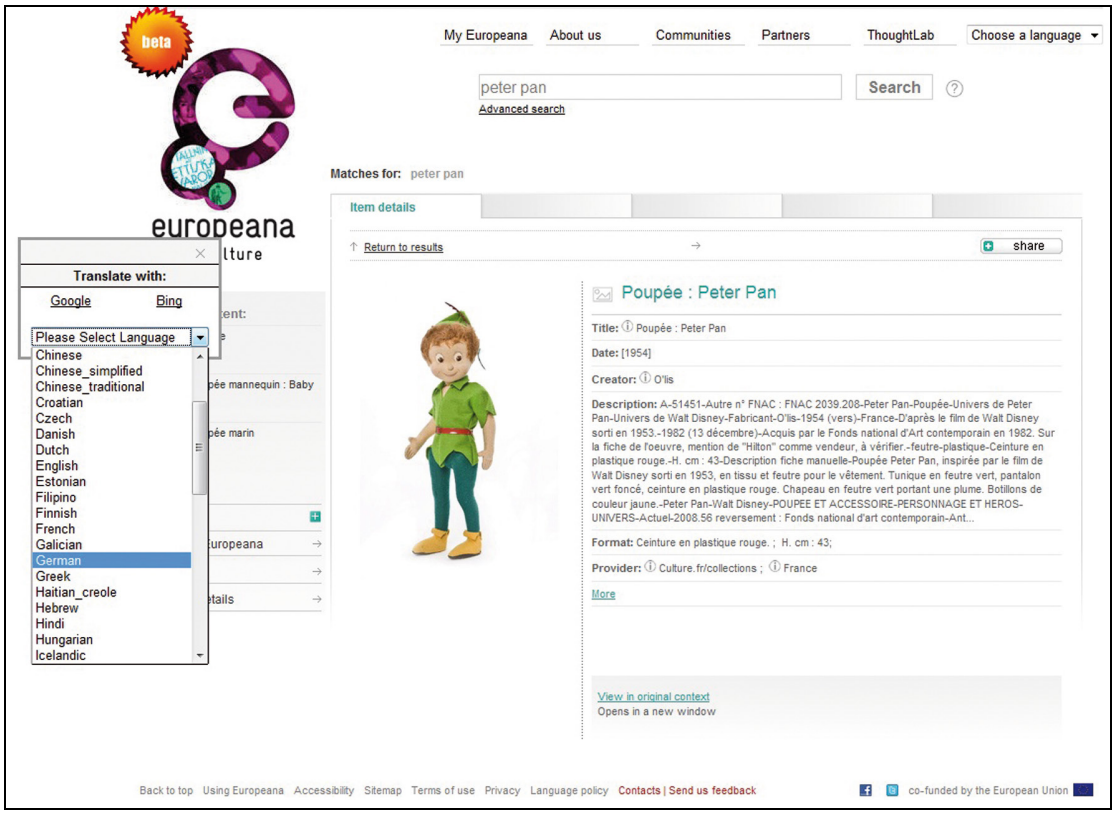

Figure 3.4 Result translation for full object view via Google and Bing services

Table 3.1 summarizes to what extent MLIA components were integrated into the Europeana portal during the data collection period. While Europeana provides multilingual interface and result representation, so far only monolingual search is possible.

Components	Europeana
Multilingual interface	Interface provided in 29 languages.
Multilingual Search and Browsing	---
Multilingual Result Representation	Result filtering through country and language facet, result translation via external services.

Table 3.1 Implementation of MLIA components in Europeana

At the time of this writing, Europeana's implementation of MLIA has not changed much. Up to date, several interface changes have been conducted but no significant changes were undertaken with regard to multilingual features.

Other dimensions of multilinguality in digital libraries are the content and user language. Europeana offers access to culturally and linguistically diverse cultural heritage objects and serves international users.

3.3 MULTILINGUAL CONTENT

The content distribution of countries and languages plays an important role when analyzing language preferences within multilingual systems. Europeana allows access to European cultural heritage objects across all languages. Depending on the content provider and metadata description, objects occur in several languages from different countries. In 2011, the Cultural Heritage in CLEF (CHiC)[26] lab was launched, focusing on the evaluation of cultural heritage information systems using test collections and queries from Europeana. For the CHiC 2012 lab, the Europeana index was grouped by country and language of the provider (Ferro et al., 2012). This categorization means that the collection language assigned according to the provider does not necessarily correlate with the language of the metadata. However, the collection language and country of origin is a useful source to investigate preferences for native language and country results.

Since the investigation of content is based on index data from 2011, it can be assumed that the distribution is similar or almost identical to the data collection period's situation. After normalization processes of inconsistent language and country codes, 29 different language collections were distinguished (Ferro et al., 2012). Table 3.2 presents the distribution of objects for the 10 most frequent languages. Almost 17% of all objects contain German metadata descriptions, followed by French (16%) and Multilingual (11%). "Multilingual" content comes from providers that aggregate objects with different languages such as The European Library.

[26] http://www.promise-noe.eu/chic-2012/home

Language	Number of Documents	Percentage
German	3,865,680	16.59%
French	3,635,388	15.60%
Multilingual	2,467,179	10.59%
Swedish	2,360,050	10.13%
Italian	2,120,059	9.10%
Spanish	1,953,124	8.38%
Norwegian	1,557,820	6.69%
Dutch	1,251,027	5.37%
English	1,107,176	4.75%
Polish	1,093,705	4.69%
Other	1,889,705	8.11%
Total	**23,300,932**	**100%**

Table 3.2 Distribution of objects per language

Table 3.3 shows the 10 most frequent countries for content providers. The determination of content per country shows a slightly different distribution compared to the language one. Sometimes, content providers offer objects in more than one language, e.g. The European Library (based in the Netherlands). Another reason is the occurrence of languages in several countries. For example, German content can appear in collections from Germany, Austria, Switzerland or Belgium. While 17% of all objects are German, only 15% are from German content providers. Most objects come from France (16%), Germany (15%) and Sweden (10%).

Country	Number of Objects	Percentage
France	3,689,138	15.83%
Germany	3,441,751	14.77%
Sweden	2,360,049	10.13%
Italy	2,120,059	9.10%
Spain	1,928,475	8.28%
Norway	1,557,820	6.69%
Netherlands	1,224,613	5.26%
Great Britain	1,096,056	4.71%
Poland	1,093,705	4.69%
Ireland	951,203	4.08%
Other	3,838,063	16.46%
Total	**23,300,932**	**100%**

Table 3.3 Distribution of objects per country

3.4 USERS AND INTERACTIONS

Because of its mission and its content, Europeana serves users from all over the world. Within the context of Europeana, five end user groups have been identified (Dekkers et al., 2009):

- General user (is interested in cultural heritage using Europeana in addition to general search engines such as Google or Wikipedia),
- Schoolchild (visits Europeana in the context of education with a special task in mind; expects an intuitive interface and help functions with regard to search and system functionalities),
- Academic user (students and teachers expecting reliable content with context information easy to export and reuse for educational purposes),
- Expert researcher (comes with a specific information need and advanced search skills),
- Professional user (is located in the cultural heritage sector including librarians and archivists; is willing to test and improve content and services).

The different countries of origin or locations and languages of international users add another layer to the variety of user groups. Within the EuropeanaConnect project, representative

personas were developed including information about personal interests, media use, search strategies as well as motivations and needs for a special user group (Guldbæk Rasmussen et al., 2010). Language skills is an important factor for search literacy, but was not explicitly integrated into the personas characteristics.

Europeana also conducted two online surveys in 2009 (IRN Research, 2009) and 2011 (IRN Research, 2011) organized by the independent research agency IRN Research. The second survey included questions concerning the users' native language and language skills and was offered in six languages. On average, respondents had language skills in at least 1.5 other languages and 71% of all non-English native speakers could access and interact with websites in English (IRN Research, 2011). In addition, interactions from log data were analyzed with regard to general access statistics (Clark et al., 2011, CIBER Research Ltd, 2013). With respect to international users, Clark et al. (2011) found that most users were coming from France (16%), followed by Germany (14%), USA (10%) and Poland as well as Spain (each 7%).

Depending on the user's background and in particular their country and language context, interaction patterns require different levels of language support. In contrast to web search, interaction patterns in digital libraries are less explored. A review of projects within this domain identified the following interaction patterns in digital libraries (Frieseke et al., 2011):

- Search (with query input),
- Explore and discover (browsing related interactions, including the goal to discover available content without a specific information need),
- Engage (user community activities such as tagging, sharing or annotating).

Known-item searches for a specific item are precise queries aiming to fulfill a predefined information need such as: "Romeo and Juliet", "Shakespeare's work" or "Mona Lisa". In contrast, overview searches have the goal to find a variety of objects related to a topic such as "The Second World War", "country music" or "still life" *(*Shneiderman and Plaisant, 2010, p. 534). Especially for cross-language searches users need extensive support when formulating queries in foreign languages or selecting appropriate translation candidates.

Explore and discover related interactions are characterized by navigational interactions with the goal to fulfill a broad information need or simply to discover the available content without the need to formulate a search query (White and Drucker, 2007). Common ways to explore digital library content are entities or facets as access points or refinement options. Users might also visit a digital library without an information need for entertainment purposes. Multilingual browsing is an essential feature for users who do not feel comfortable searching in foreign

languages or want to discover unknown content and context. Particular challenges for the translation of classifications or other category systems are culturally diverse concepts and representations (Soergel, 1997).

Digital libraries encourage users to engage with and enrich the available content by personalization functionalities such as user profiles or tagging and annotation tools. Engage functionalities support users with the goal to improve search experience through personalization options, contribute and share user-generated content or enrich existing objects (Smith, 2008). Engage interactions are still underrepresented in digital libraries (Stiller, 2012). However, inherently multilingual user-generated content can be leveraged to support cross-lingual search tasks (Stiller, 2011).

3.5 SUMMARY

The Europeana portal provides a multilingual interface as well as multilingual content and options to filter results by language and country facets. So far, only monolingual search for multilingual objects is possible. Nevertheless, its international users come with different country and language backgrounds.

The variety of countries accessing Europeana and the variety of multilingual objects highlights its exemplary status as a multilingual digital library. Results derived from aggregated Europeana usage data may also - to a certain extent - be applied to information systems of individual Europeana content providers because both the content and users overlap.

This dissertation will undertake a detailed analysis of international sessions focusing on country and language specific interaction patterns (chapters 6 and 7). Similar to a previous log file study (Clark et al., 2011), most accesses originated from France (28.86%) and Germany (12.53%). Italian sessions (7.11%) appeared in third rank followed by Poland (6.53%) and Spain (6.32%). From the 21 observed countries, only 7 showed more than 5% of all sessions within the dataset.

The analysis of country and language level differences is conducted using usage data provided by log files. The following chapter provides an overview of previous log file studies as well as a methodological description of log file analysis as a method to study user interactions.

4. LOG FILE ANALYSIS AS A METHOD FOR STUDYING USER INTERACTIONS

Digital library studies exploit a variety of methods, ranging from system to user centered approaches. This chapter briefly introduces the methodological spectrum applied within this field and highlights log file analysis as an appropriate method to study user interactions with regard to country and language level differences. Previous studies exploiting log file data investigating user behavior in web search as well as digital libraries are presented.

Based on the research foundation, basic methodological aspects of server and client sided logging approaches are presented. Finally, advantages and limitations including data privacy issues of log file analysis are discussed.

4.1 FROM SYSTEM TO USER CENTERED RESEARCH

Studies concerned with digital libraries can be classified with regard to their goal and applied methods. According to their objectives and goals, Saracevic (2000, 2004) classified previous digital library studies in seven main classes:

- System centered (focus on technical issues like performance, effectiveness and efficiency),
- Human centered (focus on user needs and expectations through the analysis of behavior, usage and information needs),
- Usability centered (focus on the evaluation of features or functionalities),
- Ethnographic (focus on the influence of the users' cultural or community background),
- Anthropological (focus on the relationship between a system and stakeholders),
- Sociological (focus on social situation and background of user groups with regard to a specific digital library), and
- Economic (focus on cost, maintenance and other economic issues within a digital library project).

In practice, most research projects apply a variety of methods. Often, approaches are categorized by having a system or user aspect orientation.

Similarly, Kelly (2009) classifies information system studies ranging from system to human focused approaches adding the field of interactive information retrieval (IIR) in between (figure 4.1).

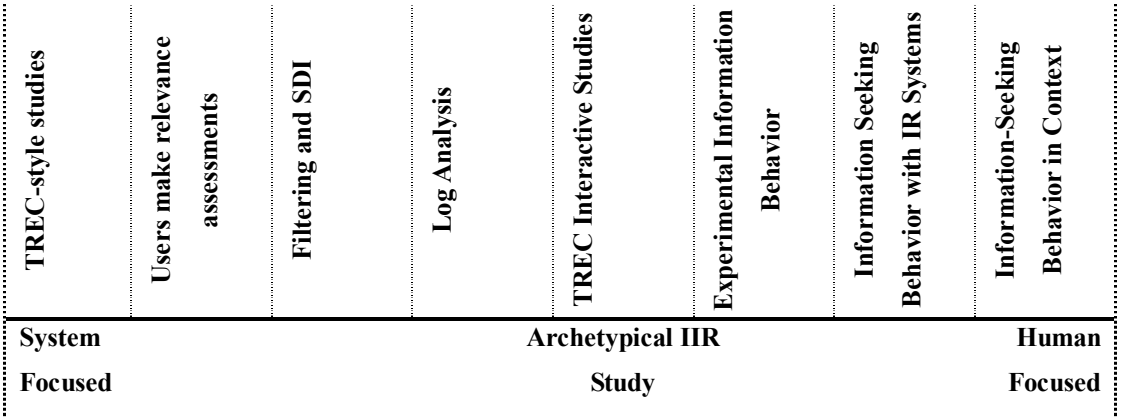

Figure 4.1 Spectrum of research studies (Kelly, 2009, p. 10)

System focused studies mainly investigate performance and retrieval metrics only involving users for topic creation and relevance assessments. Traditional system centered information retrieval studies apply the Cranfield Paradigm (Cleverdon, 1970). Cranfield-style experiments consist of a document collection, topics which are translated into search queries and relevance assessments that imply which documents are relevant to the information needs. This highly standardized test scenario has been used by many researchers and offers comparable and controlled results. While single evaluation studies are often limited to a specific system or data set, initiatives like the Text Retrieval Conference (TREC)[27], the Initiative for the Evaluation of XML Retrieval (INEX)[28], the NII Test Collection for IR Systems project (NTCIR)[29] and the Forum for Information Retrieval Evaluation (FIRE)[30] as well as the European Cross-Language Evaluation Forum (CLEF)[31] seek to move towards large-scale evaluation procedures for valuable and comparative evaluation data. Nevertheless, this approach has been criticized because of the missing integration of individual users and tasks (see Kamps et al., 2009 for an example).

While the majority of digital library studies focused on information retrieval metrics, more recent research emphasizes the importance of a more user centered or cognitive viewpoint on information retrieval research (Ingwersen and Järvelin, 2005; Kani-Zahibi et al., 2006; Saracevic, 2004; Borgman and Rasmussen, 2005; Chowdhurry et al., 2006; Khoo et al., 2008; Dobreva et al., 2012; Nicholas and Hungtington, 2010; Xie, 2006, 2008).

[27] http://trec.nist.gov/
[28] https://inex.mmci.uni-saarland.de/about.html
[29] http://research.nii.ac.jp/ntcir/index-en.html
[30] http://www.isical.ac.in/~clia/
[31] http://www.clef-initiative.eu//

Initiatives like the DELOS - Network of Excellence on Digital Libraries[32] investigated digital library projects observing a movement from strict system-centered content storage to user centered interactive systems (Fuhr et al., 2001; Franklin et al., 2009). The DELOS evaluation and interaction Triptych model addresses users' interaction issues (Fuhr et al., 2001, 2007). Similarly, the Streams, Structures, Spaces, Scenarios and Societies (5S) Framework for Digital Libraries extended their digital library model and evaluation framework (Goncalves et al., 2004) with user centered criteria like information quality, usability, usefulness and user background (Tsakonas et al., 2004).

User centric studies focus on the identification of user needs and expectations as well as realistic use cases as a crucial part of system development and design (Oard, 1997). Interactive information retrieval (IIR) or human computer interaction (HCI) research adds the user dimension to traditional information retrieval tasks, evaluating if systems support users in finding relevant objects (Kelly, 2009). Questions investigating the impact between users and systems are the central issue of interactive information retrieval studies (Marchionini et al., 2003).

Studies within the human computer interaction field use a variety of methods (Wang, 1999). Frequently used methods for digital library studies and their appropriate application are reviewed in Covey (2002) and Bryan-Kinns and Blandford (2000). Studies mainly rely on methods such as diary studies, questionnaires, usability testing, focus groups, interviews and log file analysis.

With a focus on multimedia and multilingual issues, the Cross-Language Evaluation Forum (CLEF)[33] initiative provides a platform and infrastructure for the development and evaluation of multilingual and multimedia information systems. Within CLEF, the first interactive retrieval track focusing on multilingual access was launched. The iCLEF lab mainly addressed search assistance including query formulation, reformulation and translation issues as well as document inspection and selection (Peinado et al. 2008; Gonzalo et al. 2008).

In 2011, the CHiC (Cultural Heritage in CLEF) evaluation lab[34] was launched, moving towards a systematic and large-scale evaluation of cultural heritage digital libraries (Gäde et al., 2011; Petras et al., 2012; Petras et al., 2013). In 2013, the interactive task focused on user interactions and experience using Europeana data (Toms and Hall, 2013).

[32] http://www.delos.info/
[33] http://www.clef-initiative.eu//
[34] http://www.promise-noe.eu/unlocking-culture

The LogCLEF track was launched in 2009 with the aim to study user behavior in multilingual search systems through the analysis of activities and search queries. In 2009 and 2010, log files from different providers were evaluated intending to analyze and classify user queries in order to understand search behavior in multilingual contexts and to improve search systems (Mandl et al., 2010,a 2010b). The dissertation draws upon this work, exploiting a descriptive unobtrusive research approach classifying sessions with regard to country level differences.

The choice of method to study user interactions depends on several factors such as research goals, focus of study and contextual circumstances. The investigation of country and language specific interactions within multilingual digital libraries tries to identify usage trends rather than individual differences of single users. Leveraging usage data provided in log files allows studying international users in their natural environment.

In this dissertation, a descriptive log file analysis approach is applied, combining user, system and content aspects with regard to country and language level differences. The next sections introduce log file studies as a method to study user interactions and provide a methodological foundation.

4.2 LOG FILE STUDIES

The analysis of website traffic plays an important role in a variety of research fields and domains (Taksa et al., 2009). Log file data has also become a rich data source in human computer interaction research.

From a behavioral science perspective, log files are traces of user behavior. However, log file studies use the term behavior in its narrower sense describing performed and observable actions. In information science, the concept of behavior does not only include actions but also user and contextual information such as an information need. Therefore, this dissertation uses the concept of interactions instead of behaviors provided by log files as "an electronic record of interactions that have been occurred during a search session between a search-based system and users searching for information" (Jansen, 2009, p. 2).

The Web Analytics Association (WAA)[35] defines traffic analysis as "the measurement, collection, analysis and reporting of internet data for the purpose of understanding and optimizing Web usage" (Burby and Brown, 2007). From a commercial perspective, log file

[35] http://www.digitalanalyticsassociation.org/

studies are often applied to commercial websites with regard to sales or navigational optimization. From a user perspective, log file analysis enables unobtrusive remote investigation of interactions between users and a particular system without the need of laboratory testing (Jansen, 2009, p. 2). Studying user interactions through log files comes with two main advantages:

- Scale: in contrast to qualitative approaches and laboratory studies, log files enable researchers to collect data irrespective of duration, scope and location. Theoretically, log files are able to capture all users of a system in their natural environment at any time.

- Power: the large size of usage data collected in log files allows identifying significant relations or differences with respect to interaction patterns.

Depending on the research question and goal of study, the analysis either focuses on single aspects such as search queries or takes complete user traces into account trying to identify patterns. Especially in web search, query analysis plays an important role understanding user motives and information needs.

Following, results from previous log file studies focusing on search log analysis or on the examination of complete user sessions are presented.

4.2.1 QUERY LEVEL STUDIES

The information need expressed in a query is a crucial unit for the analysis of user behavior in information systems (Jansen, 2006). Query log or search log analysis focuses on the identification of information needs expressed in user queries. Search logs have been analyzed either on a syntactic or on the semantic level with regard to:

- User goals (Rose and Livenson, 2004; Kellar, et al, 2006; Lee et al, 2005),
- Intentions or information needs (Broder, 2002; Jansen et al., 2000),
- Query categories (Jansen et al., 2000; Beitzel et al., 2007, 2007a; Gravano et al., 2003; Ozmutlu et al., 2002; Spink et al, 2002; Silverstein et al., 1999; Strohmaier and Kröll, 2012), or
- Query reformulation patterns (Spink et al., 2000; Lau and Horvitz, 1999).

Other studies focused on the relation between search terms and clicked results to investigate success rates, ranking and presentation of results (Fagni et al., 2006; Joachims, 2002).

Extensive research on query classification and web search trends was conducted using log files from popular search engines like Excite (Ozmutlu et al., 2002; Spink et al, 2002), Altavista (Silverstein et al., 1999), or AOL (Strohmaier and Kröll, 2012). Some of them made use of predefined classifications while others developed their categories based on the observed query content.

Other studies focused on the query level with the aim to identify information needs and user goals related to certain queries. The most popular and common taxonomy of web queries distinguishes three main categories of intention: informational, navigational and transactional (Broder, 2002). Follow-up studies applied and extended Broder's taxonomy, finding different distributions of user goals. The comparison of queries from 7 search engines showed that a majority of queries (80%) express informational goals (Jansen et al., 2007). Rose and Levinson (2004) identified 40% non-informational user goals. The manual categorizing of user goals is derived from a combination of queries and further session information such as results returned by the search engine as well as results clicked by the user.

Manual query analysis and annotation is a time consuming process biased by human annotator decisions. Recently, studies tended to apply automatic classification approaches. Search queries from AOL and Microsoft Research were categorized according to explicit and implicit goals (Strohmaier and Kröll, 2012). According to Strohmaier and Kröll (2012), an explicit goal contains at least 2 words and a verb. Human relevance assessors assigned around 8% of all queries with an explicit goal. With the automatic approach, 77% of those queries were detected. Baeze et al. (2006) created a test and training corpus containing 6,000 manually annotated queries. The queries were tagged according to user goals "informational", "not informational" or "ambiguous" and grouped by categories representing their topic. Additionally, the results clicked were taken into account providing vocabulary and further information for the classification process. Most queries were assigned to informational goals and categories like Entertainment, Business, Recreation, Society, Education and News. The automatic detection of informational goals also showed the best precision rate (70%) compared to the other two categories "not informational" and "ambiguous".

A few studies investigated differences between users and queries focusing on regional or language issues. Gravano et al. (2003) propose a binary classification of queries, arguing that results should be re-ranked according to either local or global information needs. Local queries like "houses Berlin" need to retrieve websites located or referring to the city Berlin or at least to Germany. In contrast, global queries like "Christmas" are usually location-independent. The comparison of queries from a predominantly American (Excite) and European (AlltheWeb.com) search engine showed that European users search more frequently for places and people while

American users tend to focus on e-commerce topics (Spink et al., 2002; Jansen and Spink, 2005).

While web search queries were the focus of many studies, less has been done to investigate queries within digital libraries or more specifically the cultural heritage domain. Early research reported query characteristics similar to web search studies (Jones et al., 1998). In contrast, the comparison of web search studies and domain specific information retrieval systems identified several differences between web query characteristics and digital libraries (Silvestri, 2010; Cecarelli et al. 2012; Waller, 2009).

To reduce complexity, some studies target single collections or domains. Within LogCLEF, TEL queries were examined with regard to their language and category. Bosca and Dini (2010) investigated to what extent query logs are multilingual and contain translation pairs assuming that users search in different languages. The categorization of TEL logs showed a high amount of queries for named entities that often could not be assigned to a particular language. This poses a special challenge for the automatic detection of query languages (Stiller et al., 2010; Hofmann et al., 2009).

Single queries do not always provide sufficient information about the user's information need and can be interpreted in different ways. For example, the query "Mozart" can be related to more than one information need including the person Mozart or work by and about him. Query suggestion approaches try to overcome this limitation by semantically enriching and contextualizing queries with external sources like Wikipedia or DBpedia (Hofmann et al., 2009; Meij et al., 2009; Petras et al., 2012; Aggarwal and Buitelaar, 2012; Kürsten et al., 2012). Other studies combine query logs with click-through data capturing every user activity in order to improve web site navigation, query recommendation or result representation and ranking.

Baeza-Yates (2005) points out two main problems inherent with web query mining. First, the expression of an information need and the corresponding query is often not clear or explicit. Second, problems arise due to the individual ranking algorithms of information systems. Clicked results do not necessarily represent the optimal answer to a related information need. A slightly different ranking algorithm displaying results in other positions might influence the user's clicking behavior.

Nevertheless, the more information is included into the analysis, the better one can interpret user interactions. The analysis of query streams and related clicks or actions is referred to as session or click stream analysis focusing on user paths (Joachims, 2002; Bucklin, 2002).

4.2.2 SESSION LEVEL STUDIES

Due to the stateless character of server interactions, every request is stored in chronological order independent of the requesting user or system. In order to analyze session-based metrics, it is necessary to reconstruct sessions by identifying entries from one user during one session.

Previous studies have applied different session identification methods and definitions (Lei and Ghorbani, 2004; Jansen et al., 2007a, Meiss et al., 2009). Depending on the available data, web sites identify users through IP addresses as well as the assignment of cookies and user or session IDs. At this time, two main approaches are predominantly used to reconstruct single user activities into a session sequence:

- Time-based heuristics: chronological reconstruction of entries per user with an inactivity break after a given timeout threshold,
- Navigational-based heuristics: sequences of referrer-request pairs.

Time-based session reconstruction usually follows arbitrary decisions about duration between page views or the entire session length. Previous research has criticized a single cut-off value for page view sequences since the approach assumes that users tend to spend the same time period at each page. The selected time frame for sessions varies from study to study ranging from a few minutes to one day (Silverstein et al., 1999; Qui et al, 2005; Montgomery and Faloutsos, 2001). For web search studies, the most prevailing time frame is 30 minutes. Nevertheless, the predetermined inactivity break does not necessarily correlate with usage patterns or provide a meaningful session definition (Meiss et al., 2010). Depending on the domain, it makes sense to adapt the session length to previously observed user behavior or compare different thresholds to find the most appropriate one(s) (Huynh and Miller, 2009; Munk and Drlik, 2011; Huntington et al., 2008). Discovery-based systems might show longer sessions than purchase oriented systems. An experiment comparing results from laboratory and remote digital library sessions showed that users in their natural environment might be interrupted or distracted and are willing to continue a session after a longer period of inactivity (Greifeneder, 2012).

The navigational-based method reconstructs user paths through the mapping of referrer-request pairs (Cooley et al, 2000). The referrer link indicates from which page a user is sending a request, while the request contains information about the requested page or document. The chronologically following action would contain the requested page of the previous action as referrer link. Consequently, a user path is characterized by the sequence of referrer-request pairs. Referrer-request reconstructions are in particular prone to missing referrer links due to

errors. The reconstruction also fails if a sequence is broken due to users typing in a URL or using a bookmark that results in a null referrer.

The implementation of the different approaches as well as their combinations have been criticized by several researchers that proposed alternative or extended reconstruction methods (Nadjarbashi-Noghani and Ghorbani, 2004; Zhang and Ghorabni, 2004; Spiliopoulou et al., 2003; Huntington et al., 2008). As for most data processing issues, the exploration of an increased number of indicators leads to better results.

Formal session classifications are sometimes extended with regard to session types. Based on the domain, site structure and study, sessions can be further classified by use cases, type of entry, number of actions, navigation, query sequence or topic (Cooper, 2001; Chen et al., 1998; Huang et al., 2004; Jansen et al., 2007a; He et al., 2000, 2002; Berendt et al., 2003).

In this dissertation, a mixed approach of time and request-referrer based session reconstruction is used.

4.2.3 LOG FILE STUDIES IN DIGITAL LIBRARY RESEARCH

In the past, digital libraries have used web analytic methods to identify user groups and usage patterns. In this section, previous usage of log file metrics in digital library research is reviewed with regard to domain specific applications.

The meaningful application of web metrics in digital library research is not a trivial task. In contrast to web search engines, digital library systems offer access to (highly) structured data. Search like it is understood and evaluated for search engines is only one aspect of digital library systems (Agosti et al., 2012). Alternative browsing services like thesauri, subject headings, classifications or exhibitions add another usage dimension and require additional examination when dealing with usage data. Web metrics have originally been developed to analyze user behavior and trends in web search environments. Due to domain and system differences, web search engine metrics are not always appropriate for the interpretation of digital library usage data. Some efforts have been made to define and standardize metrics and measurements for digital library usage analysis.

Using the example of session durations, Khoo et al. (2008) suggest reconsidering and combining web metrics in digital library research. In their review of web search studies, Jansen and Pooch (2001) compare the most frequently cited search engine studies with traditional IR and OPAC studies. While the studies reported similar values for simple search strategies and

advanced search functionalities, essential differences were observed for all other variables such as session lengths (Jansen and Pooch, 2001; Silvestri, 2010).

Within the digital library environment, log analysis was first applied to the usage of OPACs (Peters, 1993). Agosti et al. (2012) provide an overview of log file studies within web search and digital library systems.

Digital library usage can be measured with regard to resources or the provided service (Franklin et al., 2009). Most case studies apply log file data to identify user groups, needs, preferences and personalization functions for digital libraries (Agosti et al, 2007a).

Other studies focus on the usage of a specific (academic) user group or the usage of a particular collection (Jones et al., 1998). With regard to measures, visitors and visits are the most popular web indicators used by cultural heritage institutions within The Netherlands (Voorbij, 2010).

Covey (2002) interviewed Digital Library Foundation (DLF) members inquiring which methodologies they applied so far. Log files were mainly investigated in order to identify user groups, patterns and to inform interface design as well as content development or enrichment. The DLF respondents found it most challenging to collect meaningful and purposeful data as well as the processing and analysis of this data (Covey, 2002). In her review of digital library studies, Carol Tenopir (2003) found that differences in user behavior were mainly studied on a demographic or literacy level. Most studies included subjects from an academic setting. Only very few researchers involved real end users in their surveys. Especially for studies exploiting log files, no standard procedures and metrics could be determined (Voorbij, 2010).

Several projects and campaigns have emphasized the need for standardized log formats and analysis for digital library evaluation (Klas et al., 2006; White and Kamal, 2006). Within the CLEF initiative, the LogCLEF track was launched with the objective to acquire log file samples for shared analysis of search behavior in multilingual systems like TEL. One main focus of LogCLEF was the in-depth query classification and language detection (Bosca and Dini, 2010; Stiller et al., 2010).

So far, most digital library studies are limited to general statistics and analysis tools available. The user's country or language context is only partly touched by studies, mainly analyzing the origin of sessions. A previous log file study of Europeana usage data revealed that most users were coming from France (16%), followed by Germany (14%), USA (10%) and Poland as well as Spain (each 7%) (Clark et al., 2011). The study briefly refers to preferences for native country objects for some countries but does not include any other country or language variables.

In the context of this dissertation, a logging approach as well as an appropriate log analyzer was developed as a basis for the study of country and language level differences in multilingual digital libraries (described in chapter 5). Application related information such as occurrence of facets as response to a search is not included in standard http logs. As this information can provide insights into user interactions and pathways through a system, the customized Europeana Language Logger (ELL) was implemented, delivering extended information about the user and application under investigation with a special focus on country and language issues.

4.3 METHODOLOGICAL FOUNDATIONS FOR LOG ANALYSIS

User interactions can be studied using two main data sources: client-side JavaScript tags and server-side web server logs.

Client-side approaches collect user data through JavaScript code, which is added to the source code of a website. The JavaScript page tags record every successful page request. Services like Google Analytics[36] and Piwik[37] are popular client-side solutions reducing the effort of web analytics to a minimum. The services offer standardized reports summarizing usage statistics in real time. In addition, customized analysis can be generated selecting and combining available variables.

Page tagging solutions have been applied to digital library evaluation (Lee, 2011; Fang, 2007). The use of external services comes with the advantage of external data storage and analysis without the need of internal resources. At the same time, no data control is given and analysis is limited to functionalities of the provided analysis tool.

Server-side tracking exploits server log files. Standard server-side transaction logs store every server request in a single log entry. Log files are usually written in a common log format (CLF)[38] as determined and provided by the World Wide Web Consortium (W3C). Figure 4.2 shows a log entry for the Europeana portal including information about the hostname, date and time, request page, status, requested bytes, referrer page and user agent in that order.

[36] http://www.google.com/analytics/
[37] http://piwik.org/
[38] http://www.w3.org/

```
123.123.123.123 - - [11/ Mar /2010:09:42:06 +0100] "GET/cache/image/?uri=http://
images.scran.ac.uk/rb/images / thumb /0098/00980252.jpg&size=BRIEF_DOC&type=IMAGE
HTTP /1.0" 200 2843 " http ://www.europeana.eu/portal/brief-
doc.html?start=1&view=table&query=italy" " Mozilla /5.0 ( Windows; U; Windows NT 5.1; it; rv
:1.9.2) Gecko /20100115 Firefox /3.6 (.NET CLR 3.5.30729)"
```

Figure 4.2 Apache log entry for Europeana (IP address obscured for privacy reasons)

In table 4.1, an explanation for each field of this log entry is given. The remote host field represents the client or network from which a website is accessed. This field contains typically the IP address from which the geographic location and domain can be derived. Every log entry has a date and time stamp that includes the day, month and year as well as the time of access. Through the combination of IP address and the date / time stamp, it is possible to reconstruct log entries belonging to a session from one user. Another essential part of each log entry is the referrer / request pair. The referrer link indicates from which page a user is sending a request. Referrer links can either be external pages a visitor clicks on to access a website such as search engine result pages or internal pages navigating through a particular website. The type and goal of a transaction is stored in the request link. In addition, the status and size of every request is stored. Finally, background information about the operating system and browser are delivered. In this case, the Mozilla browser was used in a Windows operating system.

Web server logs can be extended or customized depending on the required measures or statistics. For this dissertation, an extended logging format was developed to focus on country and language level indicators.

Log Entry	Example	Explanation
Remote Host	123.123.123.123	The IP address of the accessing computer.
Date	[11/ Mar /2010:09:42:06 +0100]	Date and time of entry in relation to Greenwich Mean Time.
GET/cache/image/?uri=http:// images.scran.ac.uk/rb/images/thumb /0098/00980252.jpg&size=BRIEF_DOC&type=IMAGE HTTP /1.0"	Request	Transaction type „GET“ and requested page or object.
200	Status	Resultcode: 200 = successful request, 404 = error.
2843	Bytes	Size of request.
http ://www.europeana.eu/portal/brief-doc.html?start=1&view=table&query=italy	Referrer	URL of the page the user is requesting from.
" Mozilla /5.0 (Windows; U; Windows NT 5.1; it; rv :1.9.2) Gecko /20100115 Firefox /3.6 (.NET CLR 3.5.30729)"	User Agent	Information about the operating system and browser used.

Table 4.1 Fields in a Common Log Format entry

Several commercial and free web analytics tools for server log analysis like AWStats[39] are available, providing basic usage statistics. While these tools are ideal instruments for general questions, they do not support deeper insights into user behavior.

[39] http://awstats.sourceforge.net/

The analysis of log file data contains several processing steps that can be divided into three main tasks (Jansen, 2006, 2009):

- Data collection,
- Data preparation,
- Data analysis and interpretation.

While the collection of log file data is relatively easy, the following steps are time consuming but crucial for the realization of desired outcomes. Data preparation contains tasks related to data cleaning, organizing and exclusion of redundant data. Log data does not only contain human traffic but also stores requests from web crawlers. Usually, data needs to be cleaned for crawler requests. Entries generated by bots would influence statistics about session lengths or pages views per session. A detailed discussion of each step is provided in chapter 5, describing the customized logging format and analysis used for this study.

4.4 THE STRENGTHS AND LIMITATIONS OF THE METHOD

Studying user interactions through server log files and page tags comes with advantages and disadvantages. This section provides an overview of differences between the two main data sources used as well as strengths and limitations of the automated unobtrusive logging method. Table 4.2 compares the major advantages and disadvantages for log file analysis and page tagging approaches (Clifton, 2010). One of the main disadvantages of server-sided logging approaches is the limitation to server requests. All user interactions that appear beyond server requests are invisible. Those interactions include caching, browser interactions like the back button as well as print and saving commands (Kaushik, 2007).

In general, page tags provide a more complete user path since they are able to recognize client-side events as well as caching or proxy requests. On the other hand, firewalls can defeat page tags while log files are independent of security settings. In contrast to server logs, page tags exclude web crawler traffic based on their page requests (Tan and Kumar, 2002). Search engine crawlers browse the internet for the indexing and updating of website content. While human accesses request full page displays including graphic representations, crawlers only scan textual content.

Although page tags are a less resource demanding alternative to server log file analysis, it remains a disadvantage that researchers cannot rely on the raw data. In contrast, log file analysis allows more advanced or specialized studies with usage data.

Approach	Advantages	Disadvantages
Server Logs	Tracks also non-human access such as search engine robots.	Human traffic needs to be separated from non-human traffic.
	Firewalls do not affect logging.	No event tracking (e.g. Flash, Java Script).
	Long term data can be stored and reprocessed easily.	Resources for data storage and analysis are required.
	Log format can be customized with regard to system features or focus of analysis.	Partly incomplete session storage, since no entries for proxy or caching actions are made.
	Failed page requests are stored.	
Page Tagging	Does not track non-human access such as search engine robots.	Firewalls can constrain or defeat tags.
	Tracks client-sided events (e.g. Flash).	Code errors or failed requests result in data loss.
	Almost real time data collection and processing.	No raw data available for own analysis.
	Providers like Google Analytics store and process data.	
	Entire session storage including proxy and caching requests.	

Table 4.2 Server logs vs. page tags – advantages and disadvantages (Clifton, 2010, p. 22)

The major advantage of log data is that it automatically and passively records end users in their natural environment. The method allows capturing large-scale data with a high coverage of users. However, this method of analysis has a number of challenges and limitations. Data from log files needs to be interpreted carefully and within the context of each application or domain.

Kurth (2003) classifies log file analysis limitations according to four elements:

- System factors,
- User and use cases,
- Privacy and legal issues,
- Data analysis.

The development and change of systems under discussion poses a challenge especially for long-term studies. Due to programming and content changes, it is nearly impossible to reconstruct searches or retrieved objects as they have been observed in the usage data. To validate results, researchers need to store the system and database status as it was present during the data analysis period.

Log data only stores requests to a particular server and does not collect interactions with other web sites. During a session, several pages may be cached. This reduces traffic but also the completeness of session reconstruction when server-sided data is used (Kaushik, 2007). The same is true if a user accesses other websites and navigates through external sources. In this case, only interaction fragments are stored in the log files.

Besides problems that arise with regard to session reconstruction and completeness, it is nearly impossible to identify individual users with absolute accuracy. The same user may use several IP addresses or several users can share one IP address. Users that work with multiple browsers, for example Firefox, Internet Explorer or Google Chrome, will appear as different users depending on the browser version there are using. On the other hand, if several people use the same IP address, computer or browser as is common in local networks, all interactions will be identified as one unique session. User identification through cookies comes with challenges, too. An online survey dealing with privacy concerns related to cookies has shown that 17% of the participants delete cookie data weekly, 12% monthly and 10% daily (McGann, 2005). This might especially influence results and assumptions for new or returning visitors.

Web analytics studies also need to consider privacy issues associated with the usage and re-usage of log data and in particular private data (Hawkey, 2009; Buchanan et al., 2007). Increasing user participation within social media applications like Facebook[40] or Flickr[41] have broadened the scope of personal data. The definition of privacy in the internet varies and needs

[40] http://www.facebook.com
[41] http://www.flickr.com/

to be readjusted according to technical development. Actual user behavior often differs with their perception of privacy (Jensen et. al, 2005). However, log file studies have to meet governmental rules and regulations regarding the data collection and analysis. Privacy policies regulate what kind of data is collected and for which purposes. Websites have to inform visitors about the fact that they are collecting usage data. In 2006, AOL Research released a data set containing search queries from over 650,000 users[42]. Even though user information was removed from the data set, through the search history it was possible to identify unique users. This example shows that even for anonymized data contextual information can be used to reconstruct user profiles. Query logs can contain searches for results related to user names, their family, address or other personal data. It needs to be assured that the reconstruction of single users, their behavior or preferences is impossible. The Europeana terms of use and privacy policy regulates which user information is collected and for which purposes it is used[43].

Aside from that, difficulties arise when an attempt is made to answer questions concerning the reason for certain behavior, information needs or users' motivations (Ozmutlu et al., 2009). While researchers can see what users are doing, they cannot explain why they type a certain query, look at a document or visit single pages. The limitations of log file data and the missing background information can be overcome with complementary methods or the enrichment of log data with external sources (Grimes et al., 2007). Deep log analysis (DLA) approaches either enrich log data with the help of user information or combine single activities to search patterns (Nicholas et al., 2006; Nicholas et al., 2008; Nicholas, 2009).

Other limitations are inherent in the method itself. Due to privacy and competition concerns, most log data is only available to a small group of researchers. Almost every single study applies different log data covering time frames varying from one day to years. Even though results are published, it is hard to put them into a context and almost impossible to compare. The same usage data can be analyzed at different levels using different measures. Various definitions of values and metrics and very basic descriptions of methodologies leave the door open for different interpretations.

The analysis of log data can only provide a piece of the overall picture of user behavior. Having these limitations in mind, results based on log file data only indicate a transient snapshot of interactions limited to a certain system and time frame. However, due to the scale and power of log file data, representative patterns can be identified that cannot be gained using obtrusive or qualitative methods.

[42] http://en.wikipedia.org/wiki/AOL_search_data_leak
[43] http://www.europeana.eu/portal/rights/privacy.html

For this dissertation's goal of studying country and language level differences of international users, log file analysis allows the study of a significantly large sample of the complete user population for a certain period of time observing actual and therefore representative trends in interactions.

4.5 SUMMARY

With a shift from predominantly traditional system targeted retrieval studies to a more cognitive point of view, interactive information retrieval questions are the focus for recent studies. Log file analysis as a method to study user interactions positions itself in between classical system and user centered approaches. While extensive research has focused on web search queries and sessions, fewer studies have investigated users and the usage of digital libraries. Even less research deals with the impact of the user's context on search behavior and in particular on their origin and native language. Through the application of log file data, international users can be studied within their natural environment. For the purpose of this study, a customized logging approach and analysis tool was developed that gathers extended information about the user's origin and language background as well as the system and content language. The dissertation proposes to investigate variables of interactions which can be used to create country profiles and comparing country and language level differences. These variables or indicators are visible to the systems, making them appear in log files.

5. A COUNTRY AND LANGUAGE SPECIFIC LOGGING METHOD AND ANALYSIS

Different to other user contexts such as gender, age or education, a user's location is delivered by trace data. So far, log file studies have mainly touched the surface of country and language specific usage data without leveraging all available information. In the context of this dissertation, a logging approach as well as an appropriate log analyzer were developed as a basis for the study of country and language level differences in multilingual digital libraries. Application related information such as occurrence of facets as response to a search is not included in standard http logs. As this information can provide insights into user interactions and pathways through a system, a customized logging format was developed, the Europeana Language Logger (ELL), delivering extended information about the user and application under investigation.

In this chapter, the Europeana Language Logger (ELL) and its characteristics are explained, including a description of the applied variables, selected countries, languages and statistical tests. Based on the required country and language context information, direct and indirect indicators for country and language level differences provided by log file data are identified serving as a basis for the logging approach and aiming at answering research question one:

RQ1: Which indicators in log files can be leveraged to study the user's country and language context?

A corresponding log analyzer gathers specific statistics to identify country or language specific interaction patterns.

5.1 COUNTRY AND LANGUAGE INDICATORS IN LOG FILES

The identification and selection of indicators in log files that can be leveraged to answer questions about the user's background is based on the following questions:

- Where are users coming from? What is their present location?
- What is the user's native or preferred language?
- Do users change the interface language? If yes, do they prefer their native language?

- Which languages are predominantly used for searching / browsing?
- Do users refine their search by country or language facets? If yes, do they prefer their native country or language?
- Do users prefer content from their native country or language providers?

Summarizing, the main question addresses if and to what extent country or language level differences exist and which interactions differ most.

Log file data can provide information about the user's (native) language, the browser's and website interface language, the system and content language as well as user interactions with the content. Data derived from log files is either explicit, directly delivering user or system information, or implicit with the need to be processed and / or interpreted. A lot of language information in standard log files is implicit (e.g. the user's native language), which means that the indicators are logical estimators for occurrence but errors due to biases or non-standard behavior can occur. The phrase "preferred language" will be used as a signal for the user's most commonly used language (which does not necessarily mean the native language). For this dissertation, both explicit and implicit variables were determined and gathered from Europeana log files. While these country and language related variables were derived from Europeana, they are in large part valid for most digital libraries (even if some of the features are differently named).

Table 5.1 shows examples for explicit indicators. This includes all information directly extracted from user interactions stored in log data. Strong indicators for preferred languages are interactions where the user actively changes or chooses the interface language and / or the language of results. Alternatively, users can click on a link requesting the local website version from a referrer page such as search engine results. Other sources for language information are user profile settings where users predetermine their preferred interface or result language. Finally, the Europeana portal offers a language and country facet to refine search results according to their origin[44]. Common logging approaches do not explicitly log interface languages, result languages or interface changes.

A common language preference that is maintained beyond a particular website can be observed through language versions of the browser or the operating system. Unfortunately, not every browser sends its language information.

[44] Note that both facets refer to the collection's provider and do not necessarily represent the object or metadata language or country. Objects from providers collecting content from several institutions like TEL are summarized under the facet "mul" (multilingual) and "Europe".

Explicit Indicators	Explanation
Interface language (change)	The interface language or the interface language change from the default setting to a preferred language via a drop down menu, cookie or link to the appropriate language version.
Language and country of external referrer	Language and country version of external links (e.g. from Google).
Language facet / country facet	User refines result lists according to the language / country of the collection.
User account	Users can determine their native and preferred language in a user account.
Language of the user's operating system / browser language	Information about the browsers can include the language version.
Language of results	Language of an object (e.g. documents) and / or metadata language.

Table 5.1 Explicit indicators from log files

In addition, log files contain implicit information about the user's country of origin as well as the preferred and search language (see table 5.2). An indication about the user's native language can be derived from the IP address, for example. IP addresses can be assigned to a country or even a region, which again relates to an official language of this country. Another indicator is the user input in form of queries, saved searches within the user profile as well as tags and annotations, assuming that queries and user generated content mainly appear in the user's preferred language (Stiller et al., 2011). Some profiles like the one provided by The International Children's Digital Library[45] store information about the user's country of origin as well as preferred languages.

In contrast to explicit indicators, implicit information needs to be interpreted and is therefore less reliable. For example, the correlation of IP addresses to countries does not necessarily reflect the user's background but is based on the assumption that the majority of users are located in their country of origin or the country they are currently living in. This interpretation does not always refer to the user's preferred language meaning that some users are clustered by the language of the country they are accessing a website from. Also, the correlation between countries and languages poses some challenges (Gäde et al., 2011). For several countries like Belgium, Switzerland or Luxembourg, more than one official language exists. Even though the official language can be clearly determined, this does not necessarily infer the native language of the user. For example, a German user living in France would be determined as a French

[45] http://en.childrenslibrary.org/

speaker by this logic. Nevertheless, grouping sessions from different countries helps identifying general usage patterns from the same language group.

The high amount of named entity queries in cultural heritage is often language independent or ambiguous. Therefore, the query language is only partially applicable to detect the user's (preferred) language.

Implicit Indicators	Explanation
IP address	Country of origin, official language.
User account	Language of saved searches.
User-generated content	Language of tags, annotations.
Query language	Language of search terms.

Table 5.2 Implicit indicators from log files

Based on this set of indicators, a country and language specific logging approach was developed and implemented for Europeana. The next section associates the theoretically identified indicators to available variables within Europeana log files that will be used to identify country and language level differences.

5.2 CONCEPTUALIZATION OF VARIABLES

The study considers 20 variables derived from Europeana log files with the aim to identify differences and similarities in interactions between country or language groups. The variables and results are summarized under the three components of multilingual digital libraries: interface, search and browsing and result representation. Each individual variable is described with the level and impact of analysis. For better reading, variable abbreviations are used whenever all variables are considered.

Multilingual Interface. The study analyzes and compares interface language parameters from the user agent (browser), the Google search engine as well as the Europeana portal. One hypothesis is that countries are focused on their own language, which will also have an impact on search and browsing interactions. The investigation of each interface language variable provides deeper insight into preferences and acceptance of automatic and user-assisted native interface language solutions.

- *Language of External Google Referrer (GL).* The Google search engine uses a geo-based localization approach to direct users automatically to their native language version. Referrer links from the Google search engine are investigated with regard to

language parameters finding out whether the language of the referrer correlates with the assumed preferred language(s).

- *Browser Locale Language (BL).* The browser's language version is an indicator for general language preferences. For every browser that delivered its language setting, this information is extracted and investigated if country groups prefer their native language browser settings.

- *Interface Language (UIL) & Interface Language Change (LC).* The interface language and the change from the default interface language are investigated with regard to the frequency of interface language change (types) and preferences for native language(s). The analysis investigates which countries tend to use the default interface language and which tend to switch, which language change types are predominantly used for Europeana and if users prefer their native language compared to the default English version or other available languages. The analysis provides an insight into the actual usage of a user-triggered interface language change option in comparison to the automatic approach used for Google.

Multilingual Searching and Browsing. Users can choose between different ways to access Europeana's content. Within this component, variables are investigated that characterize the search or browsing process of different country groups. These variables target the question whether countries show different preferences for search options and search intensity. The quantitative analysis is supported by a brief qualitative investigation of a sample query set.

- *External Access Point (EA).* Common external entry points are search engine result pages or other links directing to the Europeana portal. Through the examination of sessions directed from an external access point, the visibility of Europeana and its content across countries is analyzed. For external links, the language of the referrer can be examined related to the country of access. As an example, this study investigates language parameters from the Google search engine (see *Language of External Google Referrer (GL)* above*)*.

- *Bounce Rate (BR).* Bounce sessions only contain one page view without any further website interaction. Due to their shortness and lack of interaction, these sessions are particularly hard to interpret. Nevertheless, a high bounce rate for specific country / language groups can indicate problems with a system that might be specific to certain countries.

- *Login.* Users can log into the portal using the user profile MyEuropeana. Within the user profile, search terms and objects can be stored and tagged. The investigation of the profile usage aims at answering the question whether differences between countries exist with regard to personalization or customization preferences.

- *Search (SS) & Browsing Sessions (BS).* Sessions with at least one query are classified as search sessions. Browsing sessions are defined by the usage of the "people are currently thinking about" (PACTA) browsing feature. Both variables are analyzed with regard to preferences for either search or browsing options. Since Europeana is a search dominated system, a low search rate from a specific country might lead to the assumption that the portal does not meet the user requirements for a language group. In contrast, a high usage of browsing features could be an indicator for effective content representation.

- *Duration of Sessions (D) & Unique Queries per Session (Q).* For each country, the duration of a session in minutes as well as the unique number of queries per session is determined and compared. The comparison of session length and queries per sessions indicates to what extent countries interact with the portal. However, a long session with multiple queries does not necessarily correspond to successful sessions. The correlation and interpretation of both variables will be related to other usage variables such as the intensity of paging behavior.

Multilingual Result Representation. This component deals with interaction issues after a user arrived at the result page including paging behavior, the use of result refinement through the country or language facet as well as preferences for native content. According to the available content within Europeana, countries with more than 10% native language content are considered as content-rich. In contrast, countries with less than 10% native language content are classified as content-poor. Individual countries as well the groups of content-rich and content-poor countries are compared with regard to native content preferences. Countries with less native language content might show less facet refinement and content interaction. Furthermore, a switch to objects in foreign languages or content provider websites is expected.

- *Brief Result Paging (BRP) & Full Result Paging (FRP).* Digital libraries encourage users to explore their content and navigate through result pages. The paging behavior within brief and full result representations is extracted in order to determine the extent of interaction after a search was conducted. It is examined whether a difference exists with respect to result paging and full object views. Differences are expected from straightforward search paths with single object views to overview strategies being selective with regard to brief and full views.

- *Usage of Language Facet (LF) & Country Facet (CF).* For each country, the top three returned language and country facets are examined to determine the occurrence of native content within search results. In a second step, similarities and differences of language and country facets usage as well as the *Selection of Native Language Facet (NLF)* and *Native Country Facets (NCF)* are analyzed. One hypothesis to confirm would be whether users preferred their native language content if it was available.

- *Language (NLC) & Country of Collections (NCC) viewed.* The collection's language or country of origin can indicate a preference for results in the user's native language. It is investigated whether countries belonging to the content-poor groups show less native content usage.

- *Outlinks to Content Providers (OL).* Since Europeana only provides metadata, the original objects remain with the content provider. In order to view the original objects, users can follow a link directing them to the content provider website. It is determined if countries differ with respect to original objects views.

The frequency of occurrence for each variable in each country group was counted. For the country groups, the unique queries and session duration as well as the corresponding median values were computed. In chapter 6, each individual variable is discussed with its occurrence per country. Results for the pair-wise comparison of countries indicating country pair differences are provided in Appendix E.

5.3 EUROPEANA LANGUAGE LOGGER (ELL)

The Europeana Language Logger (ELL) is a customized logging approach, which tracks user paths in Europeana with a special focus on country and language information. It was developed intending to observe user groups from different countries identifying usage patterns and language preferences from their interactions with Europeana (Gäde et al., 2010). The ELL combines general log data with additional clickstream actions. Like common log files, clickstreams store a single action for every server request but additionally track information like application state changes. Providing a more detailed user path, clickstream logs try to overcome disadvantages of log files (Montgomery et al, 2004).

For each request that comes to the server, a single thread is activated. To keep track of every action and application response, the ELL code is injected into this particular thread[46]. In contrast to traditional http log files, the ELL allows to analyze application states executed in response to each user request. For example, as response to a search the system performs several steps such as:

- set start time,
- create the query object from the query parameters in the request,
- check cookie settings,
- send the query to Solr,
- parse the response,
- create the response object,
- calculate the pager,
- calculate all facet URLs,
- determine the response type (HTML, XML, JSON, etc.),
- return the query result object to the rendering templates,
- render template,
- set end time for request,
- save information from the ELL to log file,
- send response back to the user.

[46] The Europeana Language Logger is written in the Scala programming language (www.scala-lang.org).

The above listed steps illustrate the complexity of system responses to a single user request. System performance and responses are especially important when an attempt is made to compare and correlate user, system and content languages. A lot of information that is used for the analysis of country and language level differences depends on application states and are logged for the ELL:

- Facets returned,
- Number of records returned,
- Interface language cookie settings,
- Requester header information like browser or computer locale,
- Type of query,
- Previous page visited,
- Request actions.

Actions logged for the ELL can be summarized in four interaction types (a list of all actions with their description can be found in Appendix C):

User management related interactions. This involves actions connected to the user account or personalization. These actions range from logging in and out, setting account preferences and saving or removing social tags, searches or objects.

Interface language related interactions. The ELL logs information about the interface language. It traces the interface language change as well as the type of interface language change. For every session, it can be determined whether, when and what kind of interface language change occurred.

Search, browse and navigation related interactions. This class of actions represents the user path through the system containing types of search actions including query terms, search result refinement, paging behavior and result views. Search actions can be either simple or advanced search while browsing or navigation related actions show alternative entry points starting from the time line or exhibitions provided by Europeana. Static page requests such as “About us” or “Privacy Policy” are logged as well.

Result related interactions. Information about the retrieved objects, e.g. number of results, returned facets, result refinement and the contribution of countries or languages are stored for each search.

In comparison to traditional log files, ELL log entries are bigger in size containing additional information as described above. Taking the example of a user triggered interface language change, a usual log entry would only indicate the language change by a different language parameter within the request link. For the ELL, an individual action was defined, including the information from which language to which new language the change was conducted.

Figure 5.1 shows an abbreviated example for the action "LANGUAGE_CHANGE". The log entry demonstrates a visitor from the Netherlands (country:NL), using the Dutch version of the Mozilla Firefox browser (locale:nl), changing the interface language from English to Hungarian (lang:HU, oldLang:EN). Through the definition of this action including the extended information about language change pairs, one can easily analyze user triggered interface language changes and preferences.

{ "_id" : ObjectId("4ee719d10364b61f80458dfb"),"action" : "LANGUAGE_CHANGE", "agent" : "Mozilla/5.0 (Windows; U; Windows NT 5.2; nl; rv:1.9.2.10) Gecko/20100914 Firefox/3.6.10 (.NET CLR 3.5.30729)", "country" : "NL", "date" : "2010-11-12T10:23:27.526+01:00", {...}, "lang" : "HU", "mlia" : { "locale" : "nl" }, "oldLang" : "EN"…}

Figure 5.1 Abbreviated log entry for action LANGUAGE_CHANGE

Rich log files can be leveraged investigating user-system interactions in every digital library. The developed logging format can be easily customized according to the characteristics and requirements of other digital library systems.

5.4 DATA COLLECTION AND PROCESSING

The study of Europeana logs is conducted in a three step process: data collection, data processing, and data analysis. Analyzing real systems and their users presents the challenge of an almost uncontrolled data collection environment. A system like Europeana changes over time launching interface changes or adding and removing functional features. Content may be added, changed or removed. Other factors influencing the number of visits and usage patterns are system errors, broken links or even server breakdowns. Logging of end users in natural settings always needs to be analyzed and interpreted carefully and, if necessary, explained through external context information.

Data Collection. In total, a dataset of ten months of user interactions from August 2010 to May 2011 containing 100,443,908 page views was collected and stored at a restricted server for further processing steps. The same dataset was used from the beginning of this study for

iterative pre-tests. During the data gathering period, two major events took place: (1) the Europeana portal was optimized with regard to search engine crawling by the end of 2010 and (2) the portal released a new version around April 2011. For both events, an extraordinary increase of page views was recorded[47]. Table 5.3 presents the number of page views for each month (after removing crawler interactions as explained in the next sub-section).

Month / Year	Number of Page Views
August 2010	594,427
September 2010	1,158,264
October 2010	1,284,213
November 2010	1,351,888
December 2010	1,119,953
January 2011	3,790,714
February 2011	6,942,036
March 2011	8,391,254
April 2011	14,435,912
May 2011	7,177,746
total	**46,246,407**

Table 5.3 Number of page views per month

Data Processing and Enrichment. The raw log entries were imported into MongoDB[48], a document-based database, for further data analysis steps. As mentioned before, log file data contains non-human usage data. Crawler requests, for example, do not represent actual user interactions and should therefore be excluded from the analysis. Several data cleaning approaches as well as their advantages and disadvantages have been used and discussed in previous research without resulting in an agreement (Stassopoulou and Dikaiakos, 2007; Doran and Gokhale, 2011).

For this research, non-human requests are removed from the sample set through the exclusion of the most frequent crawlers determined within the dataset. Sessions generated by bots normally contain a high number of log entries for paging actions. While a lot of log studies use available lists of crawlers[49], this dissertation used a more thorough approach. User agents with more than 50 requests were manually checked with regard to crawler appearance and – when identified as

[47] According to the development of page views, the data set can also be divided into two sub-collections, representing the situation before and after the search engine optimization (SEO). The first data set from August 2010 to December 2010 contains 281,041 sessions, the second one from January 2011 to May 2011 contains three times as many sessions (964,070) for the same amount of time (five months). Nevertheless, it is assumed that the SEO does not affect language or country specific interactions and should therefore not affect this thesis' analysis. The majority of variables selected for this study describe the user path within Europeana. Potential future work with the two sub-collections could investigate the impact of SEO on external access points and landing pages. An increasing number of external access points as well as direct object views probably connected to a higher bounce rate are only two expected outcomes.

[48] http://www.mongodb.org/

[49] e.g.: http://www.robotstxt.org/db.html; http://www.user-agents.org/; http://www.useragentstring.com/pages/Crawlerlist/

a crawler – added to a list of crawlers. This ensured that for this sample and this period of time, all frequent accesses and crawler addresses were checked (as the preassembled lists could be outdated). The list of frequent crawlers used can be found in Appendix D. Although this approach like every other may not identify all non-human requests, it assures that all frequent crawler accesses in the sample are excluded from the data analysis. Due to the portal structure and restricted permissions, fewer crawler accesses were observed for entries before December 2010 / January 2011. Only after the search engine optimization, an increasing number of crawler sessions can be observed. From the total 100,443,908 page views, 54,197,501 (54%) were classified as crawler accesses, removed from the data set and saved in a separate file. The vast majority of these entries originated from the Google search engine.

Different units of analysis have been discussed in previous research. Choosing session sequences as unit of analysis requires accurate reconstruction and validation of time measurement (Burton & Walther, 2001). For this study, a mixed approach of time-based and referrer-request pair reconstruction as described in chapter 4 in combination with session cookies and time stamps from the log entries was applied. A session cookie stores every communication between a web browser and the accessed server for a given time frame. After a predefined inactivity slot, sessions are automatically closed.

In contrast to web search engine studies, session length within the digital library domain has only been investigated by a few studies. Previous studies report a longer session length than the relatively short session characteristic for web search (Khoo et al., 2008; Jansen and Pooch, 2001; Silvestri, 2010). In line with these findings, an in-depth analysis of sessions showed a high amount of session fragments that were cut off when using the standard 30-minute time frame to insert a session break due to inactivity, missing related actions performed after 30 minutes of inactivity. Therefore, sessions starting with unusual actions or page views like a direct log in or result refinements as well as a brief or full result view were extracted and a random sample of 50 sessions manually checked with regard to inactivity cut offs. Accordingly, the predetermined inactivity break of 30 minutes was extended to 60 minutes. In order to validate if the predetermined inactivity break after 60 minutes is reasonable, the field “inactivityBreak” was added to the session statistics. This added another manual check of a random test sample to validate the assumed cut-off point. The alternative time-out chosen for this study reduces the appearance of session fragments due to inactivity cut-offs.

From the 50.4 GB raw data, more than 50% were removed as non-human, incorrect or broken entries. The remaining 46,246,407 (24 GB) page views were reconstructed resulting in 1,245,111 sessions initiating from 198 countries (table 5.4).

Complete Dataset	Cleaned Dataset	Session Dataset
50.4 GB	24 GB	5.4 GB

Table 5.4 Size of datasets: complete dataset with all page views, cleaned dataset without non-human pages views and reconstructed sessions.

The user's country of origin or location and corresponding language(s) were determined through the IP address (see section 5.7). Through the translation of IP addresses, sessions were clustered in country groups. Single users cannot be traced back[50]. IP addresses are mapped to data ranges provided in geo-location databases representing a single country. Two different IP to country databases were tested with regard to country detection accuracy[51]. A random sample of 50 IP addresses was extracted and manually inserted into online available IP look up services[52] and compared to the automatic IP to country assignment. The open source database used for this study showed about 98% accuracy on country assignment[53]. Mistakes or errors may appear due to dynamic IP addresses as well as proxy servers.

The IP address provides information about the country of access while interface language settings like the browser locale, the language parameters of external links as well as the Europeana interface language indicate language preferences independent of the country of origin. Due to the low number of sessions with an interface language change, this parameter was not taken into account when identifying country groups. Similarly, external links as well as browser settings do not always contain language parameters and are therefore less applicable than IP addresses.

For this study, country affiliation as indicated by the IP address was chosen as the most reliable indicator to associate sessions in country groups. According to the ISO-3166 Country Codes and ISO-639 Language Codes, all sessions were assigned to their official country and language codes based on the IP address of the accessing user (Renard, 2007). The translation of IP addresses to countries also ensures the anonymity of sessions.

For circa 2% (24,422) of all sessions, the IP address could not be assigned to a country code provided in the country to IP database and were removed from the sample set. Furthermore, due to the high number of different country groups, it was decided to focus on countries with at least 10,000 accesses within the data set. This step removed less than 15% of all sessions. Consequently, the analysis in this dissertation is based on 86% (1,071,872) of all sessions originating from 21 countries. Interestingly, the data set also includes 4 non-European countries with frequent visits from Brazil, Canada, Russia and the United States.

[50] Due to privacy laws in Germany, single users are not subject of the study.
[51] MaxMind GeoIP: http://www.maxmind.com/; http://ip-to-country.webhosting.info/downloads/ip-to-country.csv.zip
[52] http://ip-lookup.net/
[53] http://ip-to-country.webhosting.info/node/view/9

Table 5.5 relates the number of sessions with the number of internet users per country[54]. A clear domination of French users (28.86%) can be seen, having almost twice as many visits as the following countries Germany and Italy. In general, no correlation between the number of Europeana accesses and overall internet penetration can be observed. For all countries, less than 1% of all internet users visited the Europeana portal. For France, the highest percentage of Europeana sessions (28.86%) as well as the second highest percentage of users per country accessing Europeana (0.57%) is observed. The highest percentage of internet users accessing Europeana was calculated for Belgium (0.6%) with less than 5% of all sessions. The lowest percentage of accesses compared to internet users per country was determined for Brazil (0.01%), the US (0.02%) and Canada (0.07%).

[54] Percentage of Individuals using the Internet 2000-2012, International Telecommunications Union (Geneva): http://www.itu.int/en/ITU-D/Statistics/Documents/statistics/2013/Individuals_Internet_2000-2012.xls

Country	Europeana Sessions	Percentage of all Sessions	Internet Users	Percentage of Internet Users visiting Europeana
FR	309,296	28.86%	54,473,474	0.57%
DE	134,310	12.53%	68,296,919	0.20%
IT	76,210	7.11%	35,531,527	0.21%
PL	69,984	6.53%	24,969,935	0.28%
ES	67,757	6.32%	33,870,948	0.20%
NL	58,928	5.50%	15,559,488	0.38%
US	53,537	4.99%	254,295,536	0.02%
BE	51,659	4.82%	8,559,449	0.60%
GB	49,433	4.61%	54,861,245	0.09%
CA	23,305	2.17%	29,760,764	0.07%
SE	22,267	2.08%	8,557,561	0.26%
NO	21,199	1.98%	4,471,907	0.47%
GR	20,720	1.93%	6,029,983	0.34%
PT	19,179	1.79%	6,900,134	0.28%
CH	18,814	1.76%	6,752,540	0.28%
AT	17,536	1.64%	6,657,992	0.26%
BR	13,854	1.29%	99,357,737	0.01%
RO	11,292	1.05%	10,924,252	0.10%
IE	11,102	1.04%	3,730,402	0.30%
HU	11,039	1.03%	7,170,086	0.15%
RU	10,451	0.98%	75,926,004	0.01%

Table 5.5 Sessions and internet users per country (countries with more than 10,000 sessions)

Based on the country information, offical languages were assigned to each session. Table 5.6 shows the full name, its official country code and the corresponding official language code(s) for each country. Special cases are countries having more than one official language or languages spoken in more than one country. For 4 countries (Belgium, Canada, Switzerland, and Ireland), more than one officical language was identified. For example, sessions from Belgium are characterized as "native" if either the Dutch, French or German interface language, facets or content is chosen. Based on this definition, countries with more than one official language are expected to produce more native sessions (on average) than countries with one official language. Especially for the calculation of available native content it makes a difference whether the country or language perspective is chosen (table 6.6 in section 6.3.1).

Country	Country Code	Official Language Code
AUSTRIA	AT	DE
BELGIUM	BE	NL, FR, DE
BRAZIL	BR	PT
CANADA	CA	EN, FR
SWITZERLAND	CH	DE, FR, IT, RM
GERMANY	DE	DE
SPAIN	ES	ES
FRANCE	FR	FR
UNITED KINGDOM	GB	EN
GREECE	GR	EL
HUNGARY	HU	HU
IRELAND	IE	GA, EN
ITALY	IT	IT
NETHERLANDS	NL	NL
NORWAY	NO	NO
POLAND	PL	PL
PORTUGUESE	PT	PT
ROMANIA	RO	RO
RUSSIAN FEDERATION	RU	RU
SWEDEN	SE	SV
UNITED STATES	US	EN

Table 5.6 ISO-3166 country codes and ISO-639 language codes for the 21 countries selected

After all sessions were clustered into country groups, they were further analyzed with regard to interface language preferences, usage patterns and result interactions as well as preferences for native content.

With regard to Europeana content, the country and language of returned collections representing single objects is determined matching collection identifiers to a list of content providers. The 900 retrieved collections were compared to a list of Europeana collections containing information about the provider's country of origin and corresponding language. For some collections, no entry was found due to renaming or deletion of collections.

5.5 COUNTRY AND LANGUAGE SPECIFIC LOGGING

For further analysis, reconstructed sessions were automatically enriched with statistics derived from the raw data and if necessary with external sources. Table 5.7 shows a MongoDB document representing one session entry, containing all available fields and statistics. Similar to

other log analyzer tools, general statistics such as date, time and duration of access are gathered. As mentioned above, IP addresses are translated into the corresponding country of access. In order to provide more detailed session profiles, additional (country and language-specific) parameters were defined. Sessions are structured into four fields:

- General session statistics,
- Brief result page statistics,
- Full result page statistics, and
- Interface language change statistics.

For each session, the entry point is distinguished as either external or direct referrer. A direct referrer is identified through the presence of the Europeana URL, either typed or bookmarked by the user. External referrers are links referring to an Europeana page such as search engine result pages. Since roughly 45% of all sessions and 50% of all external access points are directed from Google, it was decided to extract language and country parameters for Google search engine result page (SERP) links. Depending on the focus of analysis, additional search engines or other external referrers can be easily added to the reconstruction (see section 6.1.1).

The number of page clicks is stored as well as the related actions with their frequency and chronological order. According to the number of actions, sessions are either classified as bounce or non-bounce sessions. Bounce sessions only contain one page view without any other website interaction. Usually, bounce sessions either contain a homepage view or a direct object view directed from search engine result pages (see section 6.2.4).

Result page interactions are tracked for brief and full result pages. Both groups contain general statistics about paging interactions with additional language specific aspects. For the brief result pages, the focus lays on the usage of facets, in particular the language and country facet. Full result views are primarily analyzed with regard to the selection of native language and country content (see section 6.3.3).

The interface language change block provides information about the Europeana and browser interface language (see section 6.1). The interface language type is determined as well as the sequence of languages the user switches to. In this example, no information about the browser interface language is provided, but two interface languages are logged (English and Portuguese). A full list of all available actions can be found in Appendix C.

Field	Example	Explanation
Session Statistics		
IP	B7A5C7280ACDE2F373D32BC53B7[…]	Unique identifier replacing IP address.
Countries[55]	BR	Country code derived from IP address.
Dates	2010-09-23	Date of access.
durationInMin, start_session, end_session	4; "10:35:37", "10:40:30"	Session duration in minutes with timestamp of first and last log entry.
userHasLoggedIn	false	Session with login to user profile.
inactivityBreak	true	Session was cut off due to inactivity after 60 minutes.
pageClicks	24	No. of performed actions / page views.
bounceSession	false	Session with only one page view / action.
Actions	"REDIRECT_OUTLINK" : 4, "FULL_RESULT" : 6 […]	List of actions with frequency; here: 4 times an outlink to content providers was used, 6 times a full object views was recorded, etc.
actionOrder	"INDEXPAGE" "BRIEF_RESULT_FROM_PACTA", "RETURN_TO_RESULTS", [….]	Order of actions performed during a session; here: the session starts at the Europeana homepage where the user clicks on the browsing feature PACTA (see section 6.2.3), which directs to an object from which the user returns to the result and so on.
reqRefererPairs	http://www.google.com/search?hl=en&q=biblioteca+on+line >http://www.europeana.eu:80/portal/ ...	Request and referrer links for each log entry.
hasExternalReferrer	true	Session was started from an external referrer link.
googleLanguage, googleCountry	“en”; “com”	Language and country parameters from external Google links.
Queries	"europeana_uri:"http://www.europe	Browsing, paging and related item

[55] For the country and date field, more than one entry is possible due to dynamic IP addresses as well as sessions that started at one day and end at the next day.

	ana,eu/resolve/record/90101/07967 4BD6F20C0FB6CCAF1C6785BE6 55490406E8"" : 1, "Karl Marx" : 13, "Benjamin Franklin" : 2	queries with frequency; includes queries that are typed by the user (Karl Marx; Benjamin Franklin) as well as queries set by the system as consequence of a link request.
uniqueQueriesNr	3	No. of unique queries in session.
searchType	"Paging": 5	Search types with query frequency: „initial": first result page viewed, „paging":at least 2 result pages viewed, "facet select": first result page viewed with result refinement.
facetSelected	1	No. of search result pages with facet selection.
Statistics for Brief Result Pages		
pageViews	6	No. of result list pages.
hasPagingSessions	True	Session includes brief result paging.
uniquePagingSessionNr	2	No. of unique brief result. paging sessions per query (Karl Marx; Benjamin Franklin).
pagingSessions	"Karl Marx" : 4,"Benjamin Franklin" : 2	Queries with result page interaction; for Karl Marx 4 result pages were viewed and for the query Benjamin Franklin only 2.
usesFacetsConstraints	true	Usage of facets.
hasLangFacetSelected	true	Usage of language facet.
selectedLangFacets	"pt"	Selected language facet.
hasCountryFacetSelected	false	Usage of country facet.
selectedCountryFacets	x	Selected country facet.
hasProviderFacetSelected	false	Usage of provider facet.
selectedProviderFacets	x	Selected provider facet.
queryConstraints	"\"TYPE:\"TEXT\"\"", "\"LANGUAGE:\"pt\"\"	Content of selected facets.
countryFacet	"france", "germany", "portugal"	Top three country facets returned.
languageFacet	"fr", "de"	Top three language facets returned.
Statistics for Full Result Pages		
directFullViews	0	User landed on full view without performing a search.
uniqueFullViews	4	No. of unique full object views.
nrUniqueCollections	3	No. of unique collections viewed.

uniqueCollections	"00301", "92201", "90101"	Collection IDs.
collCountry	norway	Country of collection viewed.
collLanguage	no	Language of collection viewed.
Queries	"europeana_uri:"http://www.europeana.eu/resolve/record/90101/07967 4BD6F20C..."" : 1, "Karl Marx" : 5	Queries from which the user navigated to a full view.
Statistics for Interface Language		
Languages	"EN" : 1, "PT" : 23	For each page view, the interface language is determined; this session contains one page view with an EN and 23 page views with PT
uniqueLanguagesNr	2	Nr. of unique languages: EN, PT.
hasLanguageChange	true	Session includes an interface language change.
hasLanguageChangeFirst	false	Interface language change appeared as first action.
languageChangePairs	"EN->PT" : 1	Interface language change pairs.
userTriggeredlanguageChange	false	Session includes a user-triggered interface language change via drop-down menu.
languageChangeType	"cookie-change"	Interface language change type.
Locale	x	Browser language version.

Table 5.7 MongoDB session entry (A full list of all actions is provided in Appendix C.)

5.6 APPLIED STATISTICAL TECHNIQUES

Since no assumption about the normal distribution of the sample can be made, non-parametric tests were chosen for the analysis. Several statistical techniques were applied to test for significant differences between (a) the complete dataset and (b) between the individual countries:

1. Ward's minimum variance analysis to identify meaningful clusters of country groups,
2. Pearson's Chi-squared and the Kruskal-Wallis test to test for significant differences between all country groups, and
3. the Marascuilo procedure and the Wilcoxon-Mann-Whitney test to test for significant differences between individual countries.

Cluster analysis methods were utilized to get an impression of similarities and differences between country groups by grouping similar countries together based on their interactions observed in log file data. Clustering includes several techniques for grouping observations. One of the main challenges for cluster analysis is the determination of a useful number of groups. First, a hierarchical cluster analysis using the Ward method was applied since no assumptions could be made about the number of meaningful clusters beforehand (Ward, 1963). The Ward minimum variance method groups countries based on their characteristics providing a first impression of possible clusters. Second, the k-means algorithm estimating mean or median values for a set of predefined groups or clusters was conducted identifying high and low impact variables for country and language level differences between observations (section 6.4).

Ward's minimum variance method is an agglomerative iterative process to reduce the within-cluster variance and increase the variance between groups. The squared Euclidean distance as the sum of squared distances is the most common method to determine distances between observations. For this study, a pre-test was conducted using the squared Euclidean distance. A drawback of this measure is that the Euclidean distance fails whenever correlating variables appear. For the two digital library components multilingual search and browsing as well as multilingual result representation, strongly correlated variables were observed, showing similar values for the observed countries. Multi-collinearity of variables, i.e. several variables correlating with each other, reduces the analysis complexity and prevents identifying the impact

of single variables on the cluster analysis. Correlating variables are stronger weighted leading to misleading clustering results based on fewer unique variables.

As a consequence, the Mahalanobis distance was chosen for the analysis (Sambandam, 2003), balancing variables with high variance and highly correlating variables treating all characteristics equally (Mimmack et al., 2001). The two variables session duration and unique queries are of a different scale, requiring data standardization for the variables belonging to the multilingual search and browsing component.

Based on the Ward's clustering method using either the Euclidean or Mahalanobis distance, countries are grouped with regard to similarities within the observed variables.

To visualize the relationships and iterative clustering process, the representation through a dendrogram is helpful. Figure 5.2 shows an example cluster dendogram visualizing the results for the Ward's cluster analysis, containing three observations. A dendogram contains single "leaves" each representing one observation. Observations in the same cluster are relatively homogeneous. The y-axis presents the distance at which clusters join, representing which observations are similar to or different from each other. In this example, at height 1.0 the observations one and two join to one "clade". Roughly at height 1.7, the third cluster joins the first cluster group. In other words, observation one and two are more similar to each other than observation three.

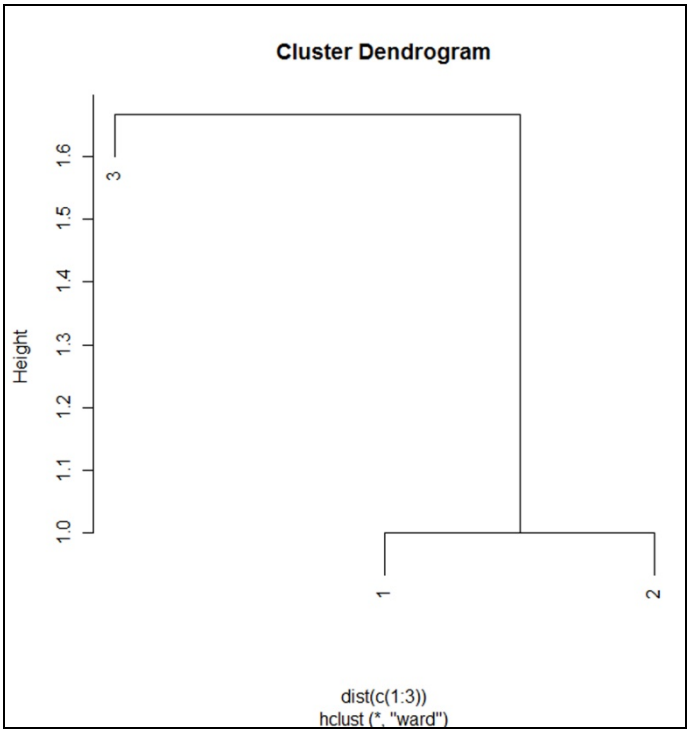

Figure 5.2 Example of dendogram visualization

Within the context of this study, countries that are joining in one "clade" are most similar based on the observed variables. The hierarchical order of "clades" needs to be interpreted from the bottom up. The higher one moves, the greater the difference between the country clusters. This means that country groups belonging to a clade at a lower height in the dendogram are more similar than clades (i.e. country groups) that are joining at a greater height.

For the *k*-means clustering algorithm, the number of cluster *(k)* needs to be determined. In this thesis, the elbow-method was applied comparing the sum of squared distances (SSD) for sequenced cluster solutions (Ketchen and Shook, 1996) based on the Ward's minimum variance method. Looking at the correlation between an increasing number of clusters and a decreasing SSD, the so-called elbow indicates at which number of clusters the SSD does not reduce substantially anymore. Selecting more clusters would minimize the variances between the clusters with no clear differences for each variable.

Using the example of multilingual interface variables, figure 5.3 displays the SSD development (y-axis) against the number of clusters (x-axis). Looking at the graph, the SSD within the cluster groups decreases significantly until the third cluster, while no significant change can be measured for the following clusters. In terms of this study, the figure suggests a division of three

cluster groups for the 21 countries observed. Although the elbow point is not always as clear as this, it offers a simple and effective way to describe a data trend. The decision of the appropriate cluster solution was again validated taking the dendograms into account. Through the application of the elbow criterion as well as the results from the Ward's hierarchical clustering, the number of clusters was determined in iterative cluster solutions optimizing *k* for the three digital library components.

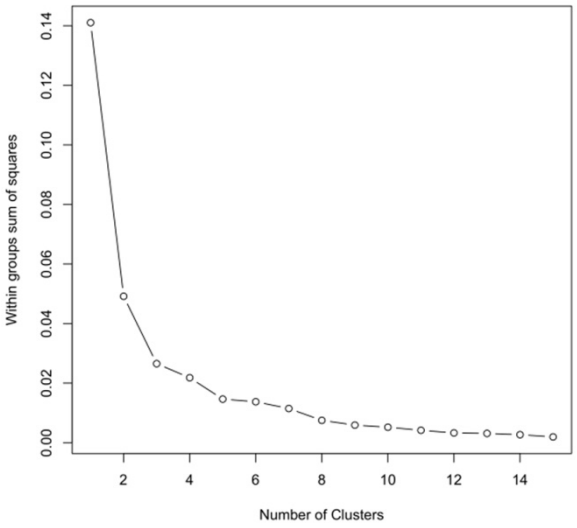

Figure 5.3 Multilingual interface variables cluster solutions

Based on the results from the k-means clustering, a ranking of variables is proposed in section 6.4. Variables with high differences between country cluster values are considered as strong variables while those that present similar values for each country are rather weak variables for showing language specific differences in user interactions.

To determine whether statistically significant differences between the countries exist, non-parametric alternatives to the t-test and ANOVA like the Chi-squared and the Kruskal-Wallis test (Berenson et al., 2012, p. 437, 457) are applied for the comparison of multiple observations. The Chi-squared test allows examining differences between more than two groups for categorical counts (true / false). For the non-categorical variables session duration and queries per session, the Kruskal-Wallis test compares if country groups show similar medians with regard to unique queries and session duration. Due to the relatively high amount of observations included, the statistical tests were expected to reject the null hypothesis (all proportions are equal), showing a highly significant difference. For all variables, this assumption was confirmed ($p=<0.001$). Nevertheless, rejecting the null hypothesis only proves that differences within the dataset exist but does not consider individual countries showing which country groups differ from each other

and which might show similar values. The computed p value refers to the comparison of all country groups. This does not necessarily mean that all countries significantly differ from each other but that at least two countries in the data set differ with respect to the investigated variable.

Pair-wise proportion tests are applied to investigate the differences between individual countries, comparing language or country groups with each other. The Marascuilo procedure enables to make comparisons between all pairs (Marascuilo, 1966; Berenson et al., 2012, p. 478-479). The Marascuilo test contains three steps. In a first step, the absolute differences between all country pairs ($c_i - c_j$) are computed. The second step identifies the corresponding critical value that determines at which point a significant difference exists. The critical range determining the boundary for significant values for each country pair is derived from Pearson's Chi-Squared test statistically adjusted for each country pair value. Finally, the absolute difference for each country pair is compared to the corresponding critical value.

Using a 0.05 level of significance, a country pair is significantly different for the investigated variable if its absolute difference is greater than the critical value. In other words, the division of the absolute and critical value is >1 for different country pairs. As an example, table 5.8 shows results for the two country pairs with the least and most difference with regard to the variable "native interface language". The absolute difference between Belgium and Germany is 0.00077 and the critical range is 0.00956. Using the Marascuilo procedure, the country pair Belgium and Germany shows a lower absolute difference than the computed critical value. Using a 0.05 level of significance, no significant difference is observed for this country pair. In contrast, France and Great Britain show a high absolute difference and a lower critical value and therefore exhibit a significant difference at level 0.05. This means that France and the GB show a significantly different behavior in selecting their native interface language whereas Belgium and Germany do not appear to be different.

Country	Country	Absolute Difference (AD)	Critical Value (CV)	Deviation AD / CV	0.05 level
BE	DE	0.00077	0.00956	0.08054393	no
FR	GB	0.84153	0.00474	177.537974	yes

Table 5.8 Example pair-wise comparison using the Marascuilo procedure

For the two non-categorical variables "unique queries" and "session duration", the Wilcoxon-Mann-Whitney (WMW) test is applied as the non-parametric alternative to the t-test. Since the WMW test compares sums of ranks of paired country differences, it is more robust with respect to outliers. Similar to the Marascuilo procedure where multiple groups are compared, the p value needs to be adjusted to control the family wise error rate (FWER) (Berenson et al., 2012,

p. 334, 494 ff.). The FWER occurs with multiple hypotheses tests, with a bigger chance for type I errors. A type I error rejects a true null hypothesis, assuming a significant difference where it does not exist. One way to restrict the FWER is the Holm adjustment (Holm, 1979). Using the Holm test, the detected p values are ordered from the smallest to the largest p value. While for the lowest adjusted value all remaining tests are considered, each following test involves one less test (Wright, 1992). This procedure avoids the detection of significance due to previous comparisons.

As an example, table 5.9 contains two country pair results for the variable unique queries. The first row shows no significant difference between Germany and Hungary at the 0.05 level (p (holm) = 0.096). Between Austria and Canada, a significant difference with regard to the number of unique queries per session exists (p (holm) = 0).

Country	Country	P Value	P Value Holm	0.05 Level
DE	HU	0.002	0.096	no
AT	CA	0	0	yes

Table 5.9 Example pair-wise comparison using the Wilcoxon-Mann-Whitney test

The pair-wise comparisons of 21 countries resulted in 210 country pairs[56]. In Appendix E, the detailed values for each variable and each country pair are provided.

5.7 SUMMARY

Log file data provides explicit and implicit country and language indicators that can be leveraged to investigate country and language specific patterns and preferences. As this information can provide insights into user interactions and pathways through a system, a customized logging format was developed. The Europeana Language Logger (ELL) traces user actions and application states as well, providing a detailed picture of interactions. A corresponding log analyzer was applied extracting and updating country and language information for each session.

A dataset from 10 months containing 1,071,872 sessions from 21 countries was determined and used for the analysis of country and language level differences with regard to 20 selected variables. Since no assumption about the normal distribution of the sample can be made, non-parametric tests were chosen for the analysis. Several statistical techniques were applied to test

[56] For the variables *Native Country Facet* and *Native Country Collection*, countries without native country options (Canada, US and Brazil) were not taken into account, resulting in 153 country pairs.

for significant differences between (a) the complete country groups and (b) between individual country groups.

In chapter 6, results from the descriptive log analysis as well as the statistical comparison of country groups are presented and discussed.

6. COUNTRY AND LANGUAGE LEVEL DIFFERENCES

This chapter presents the results from the log analysis based on the logging approach introduced in chapter 5. Based on observed trends in the specific log file sample, a careful interpretation of interactions is provided – being mindful that user intentions cannot necessarily be discerned.

Interactions clustered according to their country of origin or locations are analyzed. Country and language level differences are investigated with regard to 20 variables representing the user's context or functionalities from the three multilingual digital library components: multilingual interface, multilingual search and browsing and multilingual result representation.

Research question 2 (Does usage data indicate country or language specific interaction patterns?) will be answered by testing the two hypotheses:

- $\mathbf{H_0}$: Sessions from different countries and language backgrounds show the same interactions.
- $\mathbf{H_1}$: Country or language level differences exist between sessions.

The analysis contains three steps: First, a cluster analysis for all variables from each component was conducted with results providing a first insight into country distributions. Second, country characteristics are identified through the examination and discussion of every individual variable. Third, the analysis of each component finishes with the comparison and identification of differences and similarities between country pairs highlighting similar and different groups. With the exception of the variables "Native Country Facet" and "Native Country Collection", comparisons were made for 210 country pairs. Results for each country pair comparison are provided in Appendix E.

The chapter concludes with a classification of strong to weak variables for country and language level differences based on their cluster variances. The identification of high impact variables informs future studies within multilingual digital library research.

An aggregated consideration of all variables is presented in chapter 7, which discusses country profiles characterizing countries serving as a basis for multilingual digital library design and comparisons between country or meta groups.

6.1 MULTILINGUAL USER INTERFACE

For the multilingual user interface component, 4 variables were investigated with regard to the preferred interface language(s) for web search and within Europeana as well as the usage of the interface language change:

1. *Language of External Google Referrer,*
2. *Browser Locale Language,*
3. *Europeana Interface Language Change,*
4. *Europeana Interface Language.*

The Ward's cluster analysis for all variables showed a clear separation of English speaking countries from non-English speaking countries for the selection of a native language interface and in particular, the Europeana interface language change. As figure 6.1 illustrates, English speaking countries at the left form one cluster that differs from other country groups. Due to the fact that most systems provide English default interfaces, no effort is needed for English users to access a website in their native language. Therefore, native English sessions are characterized by a high number of native language sessions and a low percentage of interface language changes. For the remaining groups, the between cluster variance is less strong, indicating more similar characteristics with regard to the investigated variables. The dendogram shows that the countries US, CA, GB and IE are similar with respect to native interface language usage and interface language change options. Other similar country groups are BR, GR, PT, RO and HU; CH, DE, FR, AT and PL; ES, RU, BE and IT; as well as NL, NO and SE.

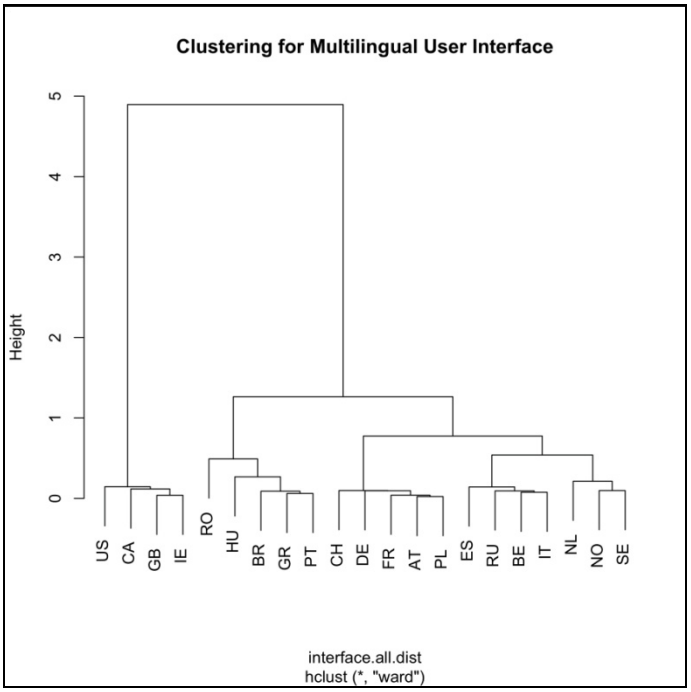

Figure 6.1 Dendogram for multilingual user interface variables

The investigation of each interface language variable below provides deeper insight into preferences and acceptance of automatic and user-assisted native interface language solutions.

6.1.1 PREFERENCE FOR NATIVE INTERFACE LANGUAGE

Browser Locale. The investigation of browser language usage is useful as it indicates general language preferences as well as the willingness of users to choose a different browser language than the default English. The hypothesis is that users choose and prefer native language browser locales or at least languages they are familiar with. All language groups of the investigated data set are represented for the most common used browsers Internet Explorer[57], Firefox[58] and Google Chrome[59]. For sessions containing information about the browser settings, the language information was extracted and investigated in terms of the users' disposition to use native browser language versions rather than other languages. The results show an average native language browser version usage of 69%. Especially the English speaking countries predominantly used their native browser versions. A stronger native language preference was also observed for Germany (87%), Poland and France (both 80%). Other countries like Romania, Netherlands, and Greece more often used non-native languages, usually English

[57] http://windows.microsoft.com/en-us/internet-explorer/downloads/ie-9/worldwide-languages
[58] http://www.mozilla.org/en-US/firefox/all/
[59] https://www.google.com/intl/en/chrome/browser/

versions. Figure 6.2 displays the values for each country with the lowest percentage of native browser locales for Romania (20%) and almost 100% native browser accesses for Canada. Canadian sessions frequently used English and French browser languages (both counted as native), while English versions appeared twice as often as French versions.

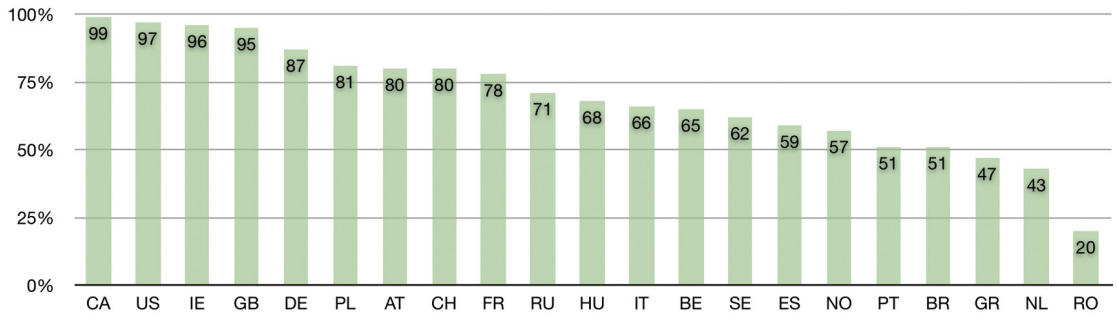

Figure 6.2 Sessions with native language browser locale

A trend is observed for frequently spoken languages to use native browser versions more often than smaller language groups.

Language of Google Referrer. External referrer links often contain language information indicating from which website version users are accessing a website. The analysis of external session entries per country is provided in section 6.2.1. Since more than half of all 1,112,223 external session entries were directed from Google, sessions starting from the Google search engine[60] (559,208 sessions) were examined with regard to native language preferences.

A clear majority of sessions (91%) were directed from the native language Google version for each country (figure 6.3). The main reason for this might be the automatic redirection based on the user's location. Nevertheless, users have the choice to select the English version as well as other language versions of Google. This might be an interesting option for users who are not located in their native country.

[60] https://www.google.com

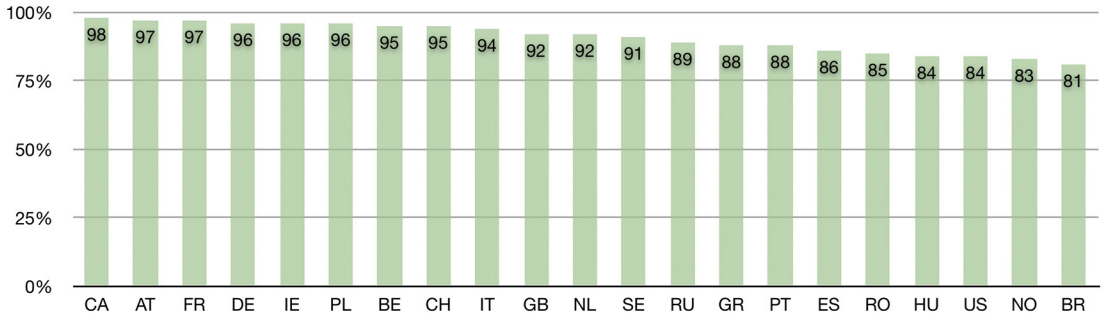

Figure 6.3 Sessions with Google native language version

While some countries show a similar native language preference as observed for browser locale settings, others do not exhibit a correlation between these two variables. For example, users from the Netherlands only used native language browser versions in 43% of all sessions, but 92% of the Google sessions were conducted from the Dutch version. In contrast, 97% of the US sessions used native browser locales but less (84%) were directed from the native Google version. With 19% non-native Google traffic, Brazil showed most interest in other languages.

For Belgium, a high preference for French and Dutch but less for German was observed while Swiss users preferred French and German to Italian. Spanish visitors frequently used the Catalan representation besides the dominant Spanish version.

Both browser locale and referrer language indicate a strong preference for native language use. In contrast, the analysis of the Europeana interface language use and interface language change indicates weaker preferences for native languages and a stronger acceptance of the default English version. It is not entirely clear whether Europeana meets the user needs for native language interfaces or should rather draw upon the experiences with general search engines like Google.

6.1.2 EUROPEANA INTERFACE LANGUAGE (CHANGE)

An interface language change from the default setting is a comparatively strong indicator for language preferences. Increasingly, studies are focusing on user interaction and acceptance of interfaces (Keegan and Cunnigham, 2005). Although it has been shown that users rarely switch the interface language (Angelaki, 2007; Agosti et al., 2007) when it is not done automatically, this variable is used to emphasize the need for localization options in surveys asking users about their needs and expectations.

During the data gathering period, Europeana offered its interface in 29 European languages with a combination of user triggered and automatic interface language change options. Three main questions are investigated related to the interface language (change) within Europeana:

- Which countries tend to use the default interface language and which tend to switch?
- Which types of interface language change are predominantly used for Europeana?
- Do users prefer their native language compared to the default English version or other available languages as the interface language?

In line with previous findings (Angelaki, 2007; Agosti et al., 2007), sessions without an interface language change - manual and automatic - appear more often than sessions with at least one language change (18%). Figure 6.4 presents the percentage of interface language changes per country. Not surprisingly, only a small minority of the English speaking countries Ireland (4%), GB (5%) and US (8%) switched the interface language. This might be an indication of non-native speakers located in foreign countries. Similarly, sessions from the Netherlands (8%), Sweden (7%) and Norway (5%) very rarely make use of the interface language change, showing a high acceptance of the default English interface. Users from Hungary (39%), Portugal (34%), Brazil (33%) and Greece (31%) show a higher use of the interface language change. In comparison, Brazil, Portugal and Brazil tail the field of countries selecting their native browser or Google language version.

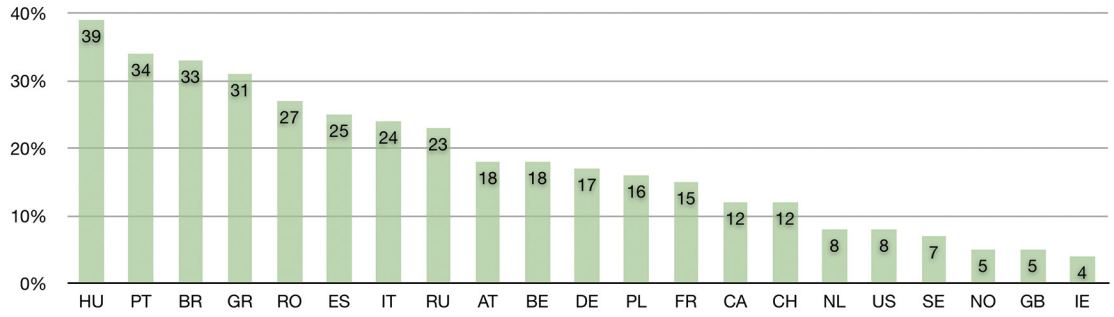

Figure 6.4 Sessions with interface language change

For those sessions including an interface language change, the frequency of language change types as described in chapter 3 is determined. The most common interface language change is the user change via drop-down menu (86%). On average, all other interface language change types as well as multiple changes like the cookie-user and link-user changes appeared in less than 10% of all sessions. Accesses from Ireland and Norway did not use links to their native language Europeana version. Most cookie changes were retrieved for Norway, Germany,

France, Poland and Italy. This leads to the assumption that users from those countries return to the portal more often.

The most frequent interface language changes are performed from the default English version to the French, German or Spanish interface (table 6.1). Interesting cases are sessions where the user switches to the same interface language, e.g. English to English (16,426) and French to French (9,959). The investigation of single sessions shows that multiple interface language changes with the same language are sometimes a combination off an initial cookie and following user change and sometimes multiple user changes. One explanation for this might be the inconsistency of the translation especially for dynamic content and the users' confusion as to what language version they are operating in. Further analysis and integration of users is needed for the interpretation of unusual events.

Language	Language	Number of Sessions
English	French	56,367
English	German	25,510
English	Spanish	18,193
English	Italian	17,144
English	English	16,426
English	Polish	10,959
English	Portuguese	10,440
French	French	9,959
English	Greece	6,945
English	Dutch	6,307

Table 6.1 Top 10 most frequent interface language change pairs

Figure 6.5 presents the percentage of sessions that switched the interface language to their native language. A clear disposition of the English speaking countries to stay with their native interface language is obvious. On average 31% of all sessions with an interface language change were executed with the assumed native interface language for each country. Leaving out sessions from the English-speaking countries GB, Ireland, USA and Canada, the percentage is even lower (16%). Interestingly, users with less frequent languages like Hungarian, Portuguese and Greek select their own interface language more often than users with frequently spoken languages like Spanish, France and German.

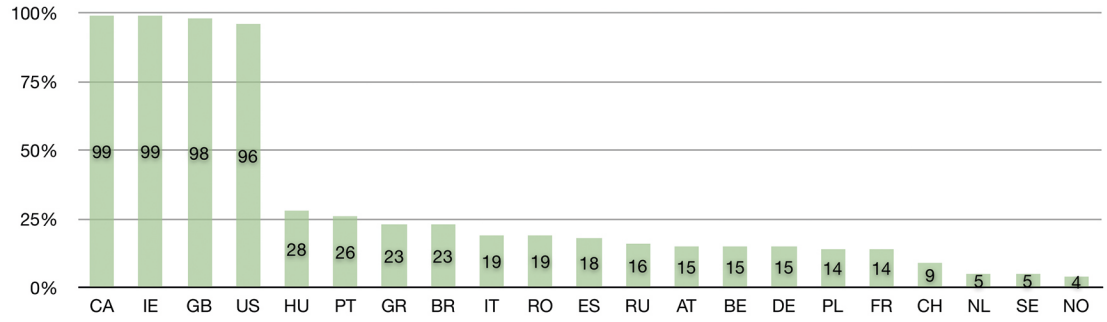

Figure 6.5 Sessions with native interface language

It is difficult to make assumptions about interface language preferences for sessions from the GB, USA, Ireland and Canada whose official language English is identical to the default settings for Europeana. A special case is Ireland, whose native language Irish was not offered as an interface language during the data gathering period.

The investigation of interface language changes within Europeana suggests that users regardless of their origin predominantly apply the default English interface version. This behavior contrasts with the frequent selection of usage of native Google and browser versions. The most common interface language change type is the user change via drop-down menu. Not surprisingly, users tend to switch to their native language rather than to other available languages.

6.1.3 COMPARISON OF COUNTRY PAIRS

It was determined that most countries make use of native language interface versions in daily life. A clear preference for native language interfaces regarding browser locale and the Google search engine was observed. In contrast, a different behavior was detected for the Europeana portal interface language. Only a minority of users switched the interface language to their assumed native language – maybe because of the infrequence of use it does not appear necessary or maybe because switching the interface language in the portal is too troublesome. Country differences are determined using the Marascuilo procedure reporting the deviation between the absolute (AD) and critical difference (CD). Results for all country comparisons can be found in Appendix E.

Statistically for 196 out of 210 country pairs the difference for the usage of the Europeana interface language change was significant. Considerable differences were calculated between non English speaking and English speaking countries. Spanish users for example switch the interface language more often and therefore differ strongly from British users (AD/CD=18.89)

who display an interface language change in less than 5% of all sessions. In line with the low number of interface language changes, British sessions mostly contain English page views in contrast to French sessions that switched to the French interface version very rarely (AD/CD=177.53).

For 93% of all country pairs (195 out of 210), significant differences were calculated with regard to native language browser versions. The least different values were observed between Portugal and Brazil (AD/CD=0.01). Roughly half of all sessions from both countries are using native language browser versions. The highest differences (AD/CD=29.98) were observed for Canada with almost 100% native language browser accesses in contrast to the Netherlands, France and Romania selecting their native language less frequently.

The usage of native language Google versions provided more similarities between country groups with 136 of 210 different pairs. Austria and France showed most similar values for (AD/CD=0.01) indicating a strong preference for their native language versions. More differences were observed for Canada with most native language Google accesses and countries like Spain that show less native language accesses (AD/CD=5.74).

6.2 MULTILINGUAL SEARCH AND BROWSING

Users can choose between different ways to access Europeana's content. Various interaction patterns are observed, due to different entry points and possibly other factors like literacy, language skills or type of information need. The following section investigates components that characterize the search or browsing process.

During the data gathering period, Europeana did not offer cross-language search functionalities but only monolingual search. Nevertheless, several other variables can give an insight into the search and browsing patterns within a multilingual system like Europeana. In total, 7 variables were investigated related to multilingual search and browsing activities:

1. *External Access Points,*
2. *Usage of User Profile,*
3. *Bounce Rate,*
4. *Search Sessions,*
5. *Browsing Sessions,*
6. *Duration of Sessions,*

7. *Unique Queries per Session.*

Exemplarily, the 100 most frequent queries from the top access and content-rich countries France and Germany are analyzed with regard to query category and language.

Compared to the visualization of multilingual interface variables, the cluster analysis illustrated in figure 6.6 suggests a stronger variance between the countries. Taking all variables into account, it stands out that in contrast to figure 6.1, English speaking countries do not longer form one cluster but spread depending on their characteristics. Especially Ireland differs from all other countries, only converging at a relatively high level when it comes to searching and browsing interactions. A similar situation is observed for the Netherlands and Poland. The cluster containing Brazil, Greece and Portugal differs highly from the other countries coming together in clusters at a lower level.

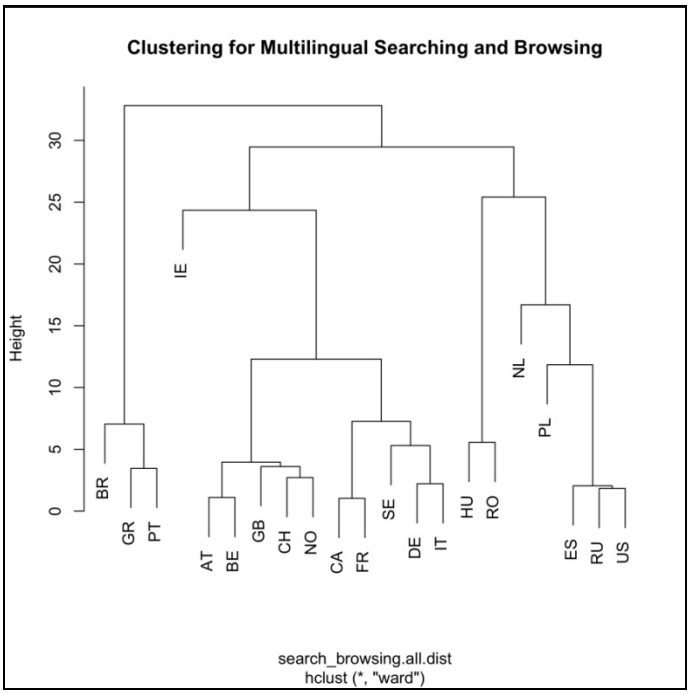

Figure 6.6 Dendogram for multilingual searching and browsing variables

6.2.1 EXTERNAL ACCESS POINTS

There are alternative ways for a user to access a website or start a session. A possible distinction would be direct or indirect access. The most common way to access a website is the indirect entry via external links such as search engine result pages. In contrast, direct access appears whenever a user types in the portal URL or loads a bookmark. The analysis of session entries provides an insight in the external website's visibility. Language parameters delivered by

external links serve as a basis for the analysis of multilingual traffic as it has been done for the Google search engine in section 6.1.1.

As figure 6.7 shows, the majority of users (89%) access Europeana via external links. In particular, users from Ireland, Switzerland and France (all 92%) make use of an external entry point. A slightly higher percentage of direct access points were observed for Hungary (15%), Greece and Russia (both 14%). All three countries belong to the group of content-poor countries (see section 6.3.1). The need for direct accesses might be a result of fewer visibility of native content within search engine results.

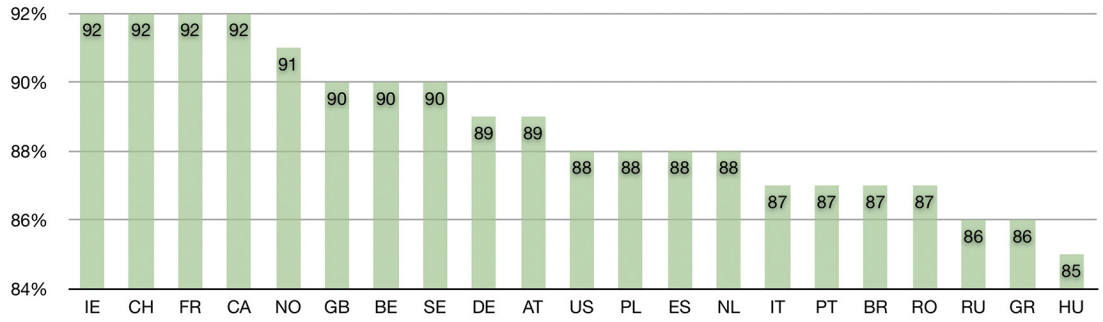

Figure 6.7 Sessions with external referrer

6.2.2 PERSONALIZATION

With regard to language preferences, the user profile is an important source for user background information such as language skills and preferences for interface language and result translation (Gonzalo et al., 2008). Only a few studies have focused on and emphasized advantages realized with personalization opportunities (Budzise-Weaver et al., 2012). The Europeana policy on user-generated content states that through multilingual user-generated content such as tags and annotations, information access can be improved (Keller and Oomen, 2010).

Europeana offers a user profile called “myEuropeana” where users can log in to save searches, objects or tags. Currently, no background information or user preferences for interface or result languages that could be used for this study as user background information are stored. The most frequent user profile usage was observed for Greece and Hungarian users (13%). Only 2-3% of all sessions from Norway, France, Sweden, Switzerland and Poland included a log in (figure 6.8).

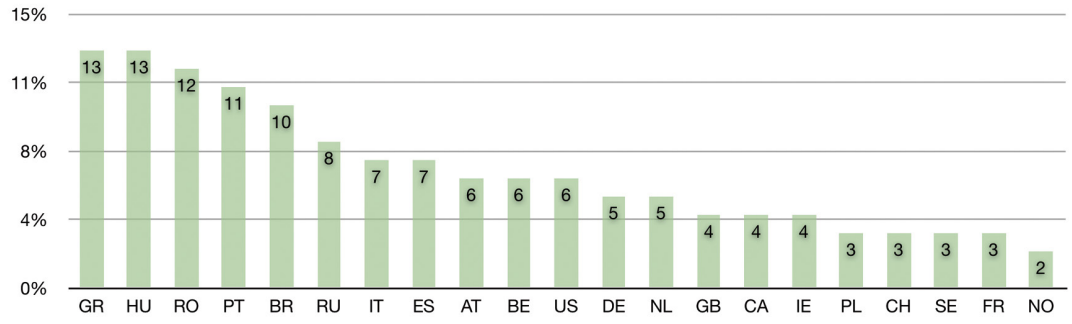

Figure 6.8 Sessions with log in

The relatively small interest in the user profile from all countries can have several reasons. One of them might be the missing aspect of collaboration and sharing options.

6.2.3 INTERACTION PATTERNS

In the cultural heritage domain, only a few studies focused on the identification and examination of interaction patterns. A review of related projects in the domain summarized the predominantly occurring interactions as follows (Frieseke et al., 2011):

- Search (traditional search box with query input),
- Explore & discover (browsing patterns including interactions with the goal to discover available content without a specific information need), and
- Engage (user activities related to content such as tagging, sharing or annotating).

According to the intensity of performance, a fourth pattern can be distinguished:

- Bounce or single page views (users that access a website without interacting at all; bounce sessions contain only one page view).

Engage actions are mainly provided within the user profile, where search terms and retrieved objects can be saved and tagged. The inspection of usage data has shown that only very few sessions contain an action related to engage patterns. Therefore it was decided to concentrate on the dominant search, browsing and bounce interactions.

Bounce. Based on the websites purpose as well as the action performed, bounce sessions can be interpreted in different ways. Bouncers are users that access a website and leave immediately

without having viewed any other page or conduct another action. Corresponding *non-Bouncers* are users that conduct at least two actions during a session.

A user searching for a particular object might be directed from a SERP to a particular landing page without the need to interact with the entire system or view other pages. Another scenario would be a user landing at the Europeana homepage page without any other interaction. According to the previous definition, both cases would be bounce sessions. Nevertheless, a user viewing a single object might be satisfied; a user only viewing the index page might not have understood the purpose or structure of the website. For system designers, it is important to identify pages with a high bounce rate. At this point, users might feel lost and do not know how to navigate through the portal.

So far, bounce patterns have only played a secondary role in investigating information seeking behavior. Nicholas et al. (2007) conducted a deep log analysis of five digital journal libraries focusing on bounce sessions. Beside other characteristics, they determined the geographical location of bounce sessions and found that Eastern Europeans tend to visit a website only once and therefore are more likely to be classified as bouncers (Nicholas et al., 2007). A previous study of Europeana logs made clear that bounce sessions (65%) are predominantly coming from search engine result pages, often viewing one particular object (Clark et al., 2011).

On average, 16% of all sessions were characterized as bounces. Figure 6.9 illustrates the percentages of bounce sessions per country. Swedish, Norwegian and French users tend to bounce more often (20-21%), while users from Brazil or Hungary showed the least bounce rates (11%). In contrast to Nicholas et al. (2007), Eastern European countries are among the weaker bounce country groups.

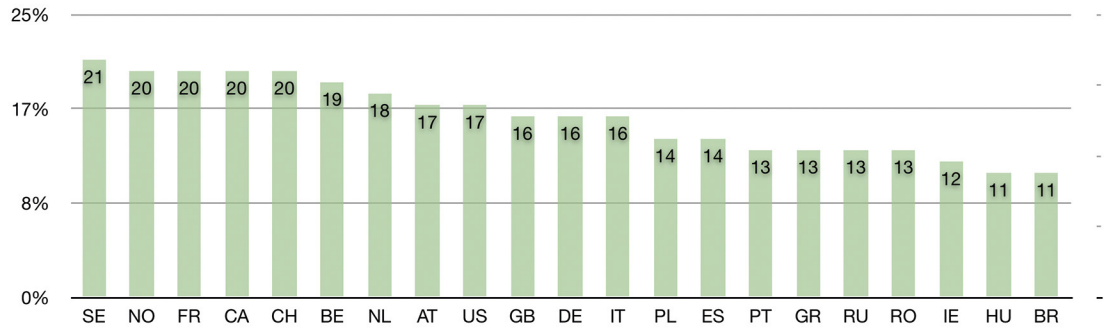

Figure 6.9 Sessions with single page view

Search and Browse. Search sessions are defined by the occurrence of at least one query per session. For all countries, sessions with at least one query (80%) dominate. The high amount of

sessions containing queries might be a result of the search dominated user interface providing a simple search box as well as advanced search functionalities. At the time of writing, the advanced search interface is no longer in use. As a compensation, users can select a metadata field within the simple search box to specific their search terms.

French users (90%) conducted at least one search most frequently, followed by Swedish (88%) and Polish users (88%). The lowest percentage of search sessions was observed for Brazil (73%), Portugal (68%) and Greece (68%) (figure 6.10). An in-depth study of non-search sessions without any query might provide an explanation for fewer searches from certain countries.

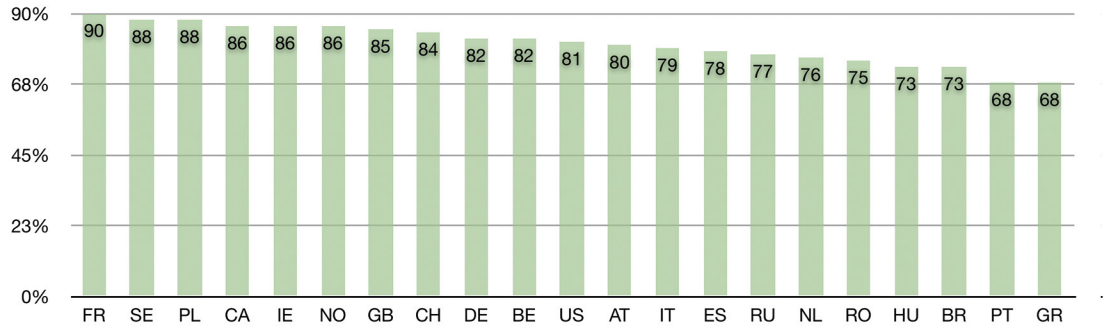

Figure 6.10 Sessions containing at least one query

Alternatives to classical search interactions are browsing related actions. Users who do not have a clear information need in mind or merely want to explore the available content make use of browsing features. Koch et al. (2004) studied activities within a browsing focused service. They observed a preference for browsing activities (80%) due to the website structure and a high number of entry points at browsing pages from SERP. The study poses the question to what extent system design influences user behavior and what makes users prefer searching or browsing activities.

During the data gathering period, Europeana was a search dominated service providing only limited browsing access such as the „People are currently thinking about“ (PACTA) feature (figure 6.11). Using previous searches, query suggestions were presented to the user at the homepage. According to the interface language, the suggested query terms were displayed in the appropriate language. Since the PACTA functionality represents the only language-sensitive browsing option, it was decided to focus on this particular interaction rather than on time line browsing.

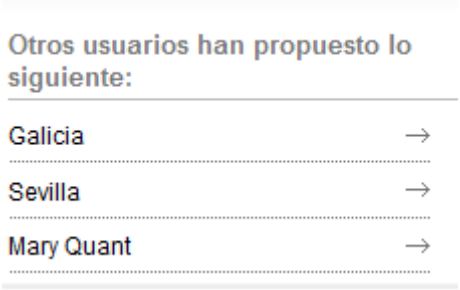

Figure 6.11 Query suggestions from Spanish interface (2011)

Irrespective of the user's origin, only a very small proportion (3%) made use of queries presented by PACTA (figure 6.12). Countries with the smallest amount of search sessions browsed more often than other countries. The majority of PACTA clicks originated in Greece (7%), Portugal (6%) and Brazil (6%). Less than 1% of all sessions from Norway and roughly 1% of the sessions from Sweden and GB clicked on a displayed search suggestion. The relatively low usage of PACTA might have several reasons varying from the search dominated interface to misunderstanding and lack of interest for this particular implementation. With the new release of the Europeana portal, this feature is no longer offered.

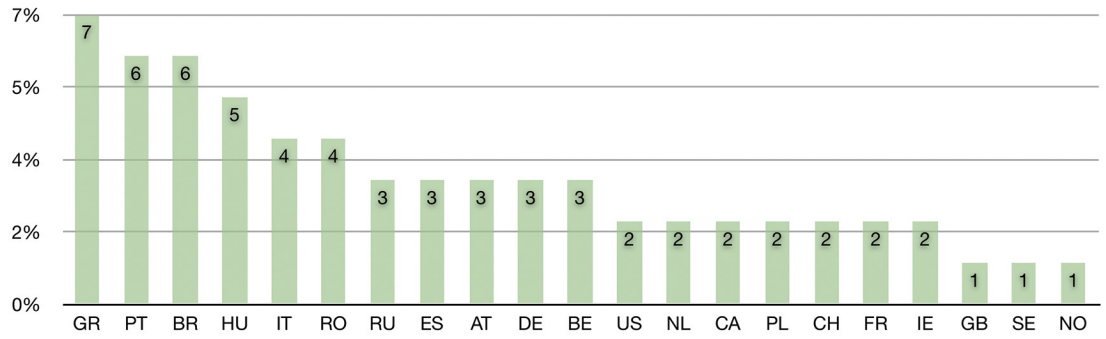

Figure 6.12 Sessions with query suggestion (PACTA) usage

6.2.4 SESSION DURATION AND UNIQUE QUERIES

With regard to the duration and number of queries per session, previous analysis of TEL and Europeana logs found that the majority of search sessions involves one query (Angelaki, 2007) and usually does not last longer than two minutes (Clark et al., 2011).

In line with previous findings, the majority of sessions last less than one minute and only contain one search query. In contrast, a few sessions contain extensive query input lasting more than 30 minutes including inactivity slots. On average, sessions contained 1.8 unique queries and lasted 24 minutes. However, the average numbers may not all be representative for the

majority of sessions showing high standard deviations due to the range from predominantly short and unusual long sessions with more queries. Table 6.2 presents the average session duration in minutes as well as the unique queries with the standard deviation for each country. The highest session duration averages were observed for Russia (39 min.), followed by the Netherlands (31 min.), Poland (26 min.) and Switzerland (26 min.). Russia also showed the highest standard deviation with 185.14 minutes. The shortest session duration averages were generated by Norway (17 min.), Canada and Great Britain (both 18 min.). Accordingly, Great Britain (1.5) and Canada (1.4) showed a lower number of unique queries per session. Irish sessions contained the least input volume. For Spanish sessions, an average number of 1.94 queries was computed with the highest SD of 5.43.

Countries	Mean Session Duration	SD	Mean Number of Queries	SD
RU	38.92	185.14	2.14	4.68
NL	30.60	167.62	1.76	4.59
GR	27.27	152.35	1.90	3.80
PL	26.28	151.10	2.18	4.03
CH	26.00	153.78	1.51	3.23
BR	25.59	147.66	1.89	3.18
BE	24.40	151.62	1.71	2.50
PT	24.38	137.06	1.70	2.90
FR	23.85	149.63	1.64	2.68
AT	23.73	148.97	1.82	4.31
ES	23.43	142.50	1.94	5.43
IT	23.03	143.20	1.89	4.01
US	22.18	144.60	1.56	2.30
DE	21.91	140.90	1.77	3.32
HU	20.75	132.00	2.06	3.71
RO	19.70	133.30	1.84	3.81
IE	19.67	138.30	1.42	1.60
SE	19.67	136.70	1.63	2.39
GB	17.81	13.000	1.49	2.35
CA	17.71	129.90	1.44	2.17
NO	16.99	125.52	1.74	2.88

Table 6.2 Mean duration in minutes and unique queries with standard deviation per country

The results indicate that the session length and the number of queries do not necessarily correlate. A correlation between a relatively long session length and number of unique queries was observed for Russia and Poland. In contrast, the Netherlands have the second longest session length but only 1.7 unique queries. Similarly, Hungarian sessions are rather short in

duration (21 min.) but contain two unique queries on average. One hypothesis might be that long sessions with less unique queries show a higher frequency of paging interactions while in relatively short sessions with more queries users might tend to reformulate or type in new queries instead of browsing through result lists.

6.2.5 QUERY ANALYSIS

The investigation of queries and underlying information needs or user goals is another important issue to understand user needs and requirements. While extensive research deals with web search queries, (multilingual) query analysis in digital libraries has only been touched by a few studies (Stiller et al., 2010; Hofmann et al., 2010).

For all 21 countries, unique queries with their counts were extracted for further research. In line with the number of sessions, countries with high access frequencies also produced the majority of unique queries. The least query input was counted for Ireland with only 3433 unique queries (table 6.3). With regard to single occurence queries, all countries showed a high percentage with more than 80% of all queries occuring only once in the data set.

Country	Unique Queries	% of single Occurrence Queries
FR	147,511	87.44%
DE	83,886	83.35%
PL	68,216	81.65%
ES	62,670	86.74%
IT	55,778	84.01%
BE	33,586	85.63%
NL	32,371	81.53%
US	29,970	86.28%
GB	19,117	85.38%
GR	18,200	84.15%
AT	14,239	85.65%
BR	13,435	86.56%
PT	15,814	84.12%
RU	12,481	84.99%
HU	11,473	85.01%
CA	9,964	85.92%
CH	9,727	88.36%
RO	9,563	84.85%
SE	9,149	84.80%
NO	7,852	83.69%
IE	3,433	82.14%

Table 6.3 Number of unique queries and single occurrence queries per country

Previous studies have shown that automatic language detection of cultural heritage queries provides biased results, classifying named entities as English queries (Stiller et al., 2010). Therefore, no satisfactory result is expected using automatic language identification for the extracted queries. The manual analysis and classification of queries from different countries is highly interesting but since various variables are considered for this study, only a small example is chosen for further analysis. Multilingual query analysis of all countries is postponed to future studies with the dataset.

As an example for the possible types of analysis, the 100 most frequent queries from the top access and content-rich countries France and Germany are analyzed with regard to query category and language. The categorization of queries is based on the classification developed by Stiller et al. (2010). Table 6.4 presents the categories with a description and examples from TEL and Europeana log data. For this study, the three categories “named entity”, “browsing” and

"topical" are distinguished. As explained in table 6.4, the category named entity contains queries searching for a person, geographic name, work title, organization, event or domain specific terms. Topical queries are not assigned to a specific category as named before but contain thematic searches as well as ISBN numbers or dates[61]. For this study, the category "browsing" was added to assign queries that were set by the system as result of a link clicked by the user.

Category	Description	Example
Person	Artist, creator, scientist, politician.	Mozart, Napoleon.
Geo	Monument, town, country.	Berlin, Waterloo.
Work title	Book, article, opera, pictures.	Mona Lisa, Bible.
Organization	Institution.	NSA.
Event	Historical.	Waterloo, Second World War.
Domain specific	Specific terms for disciplines such as biology.	Candida, testosterone.
Topical	Thematic, not assigned to a specific category, ISBN, dates.	Fussball, livres, 1987, 967-323-12343-8.
Browsing	Set by the system.	Europeana_collectionName:079* Postcard OR carte postale OR postkort OR Postkarte[62].

Table 6.4 Query categories (Stiller et al., 2010)

For the query language, it is determined whether it is a native language, English or an ambiguous query. Queries are classified as ambiguous whenever it is not possible to assign them to a specific language. Especially named entities such as "Mozart" are often language independent. Through the determination of English queries, an assumption about English as secondary search language can be made.

Country	Topical	Named Entity	Browsing Query	Native Language	Ambiguous	English
DE	13%	74%	19%	32%	55%	11%
FR	27%	74%	9%	37%	71%	7%

Table 6.5 Query category and language for top 100 German and French queries (%)

For some queries, multiple categories were considered. Table 6.5 contains the result for the 100 most frequent German and French queries. For both countries, the majorities of queries are looking for a named entity. French sessions more often include a topical search while German sessions tend to browse. With regard to the presence of native language queries, both countries

[61] ISBN numbers as well as dates for a specific event could also be classified as named entities.
[62] Multilingual query generated by the PACTA functionality

show a similar distribution with only 32 and 37 of the top 100 queries that could be clearly assigned to the appropriate native language. About 55% of all German queries are ambiguous in language, for French even more (71%) cannot be assigned to a language. German users tend to use English as secondary language slightly more often than French users.

In figures 6.13 and 6.14, German and French queries are visualized through a word cloud. In line with the results of a qualitative analysis of 100 queries, both figures show a high coverage of named entities. In comparison it stands out that only a few queries overlap, while more are country specific. German queries mainly include searches for German cities and people. Topical queries often correlate with the assumed user language. In general, German users tend towards the German expression for topical queries "Adressbuch" (address book) as well as for language dependent named entities like "gotik" (gothic) or "Berliner Mauer" (berlin wall). For named entities, German users often use the upper case version.

Figure 6.13 Word cloud with frequent German search queries (without browsing queries)

French users often used the advanced search option, specifying the query via fields, e.g. "text:coupe AND text:zeus". Named entity searches are mainly querying for persons. French sessions very rarely contain English queries.

While similarities for the query category as well as for the ambiguity of the query language exist, differences are observed with regard to the query content. Previous web search studies indicating differences between countries (Jansen and Spink, 2005) as well as the qualitative analysis of the small sample give reason to assume that country specific information needs and queries exist and need to be considered dealing with multilingual search functionalities.

Figure 6.14 Word cloud with frequent French search queries (without browsing queries)

Annotated query corpora as provided in LogCLEF (Stiller et al., 2010) are the basis of further investigations. The identification of the query language adds another language dimension and leads to additional correlations with other language indicators like the interface language, external referrer languages, browser languages and the language of results viewed.

6.2.6 COMPARISON OF COUNTRY PAIRS

In total, 7 variables were investigated, focusing on multilingual search and browsing patterns. Country groups are more similar with regard to search and browsing patterns compared to multilingual interface and result representation patterns. All groups tend to access Europeana via external access points and rather use the search box than browsing features. Very little interest is observed for the user profile.

The qualitative analysis of two sample query sets from Germany and France has shown that the majority of searches are looking for named entities that are often language independent. Within the top 100 queries, only a few queries overlap which indicates that countries have different information needs.

Country differences are determined using the Marascuilo procedure, reporting the deviation between the absolute (AD) and critical difference (CD).For the two variables duration of sessions and unique queries, country differences are calculated with the Wilicoxon-Mann-Whitney test using the Holm method (see section 5.6). Results for all country comparisons can be found in Appendix E.

The statistical comparison of country groups reveals less variance between individual countries. The least different country pairs were observed for external entry points with 136 different

country pairs and for bounce sessions with 150 out of 210 pairs. For external entry points, only a few countries showed more frequent direct referrers in contrast to the majority of countries with more than 90% external access points (e.g. Italy 87% – France 92%). With 15% bounce sessions German and Italian users behave most similar (AC/DC=0.03). A higher bounce rate from France contrasts to fewer bounces from Spain (AC/DC=7.68).

Most different country values are observed for the duration of sessions with 187 significant varying pairs out of 210 comparisons. Session length varies from 17 minutes on average for Norway to twice as much time spend at the portal for Russian sessions with 39 minutes. For the average number of queries per country, a range from 1.42 (Ireland) to 2.14 (Russia) queries per session are discovered.

Country pairs differed with regard to the usage of "MyEuropeana" and the browsing feature PACTA in 170 and 171 out of 210 cases. With roughly 4% PACTA usage, Italian sessions differ significantly from Norwegian sessions with less than 1% PACTA usage (AC/DC=6.83). With regard to user profile usage, Italian users log into the user profile more frequently in contrast to less than 5% of French sessions making use of the personalization (AC/DC=8.42).

6.3 MULTILINGUAL RESULT REPRESENTATION

The following section deals with interaction issues after a user arrived at the result page including paging behavior, the use of result refinement through the country or language facet as well as preferences for native content. According to the available content within Europeana, countries are considered either content-rich or content-poor. For this study, countries are considered to be content-poor if they feature less than 10% native language content while content-rich countries hold more than 10% native content within Europeana. The classification into one of these two groups plays an important role when interpreting interactions and preferences with native language content. It will be investigated if and to what extent the availability of native content has an impact on result interaction patterns.

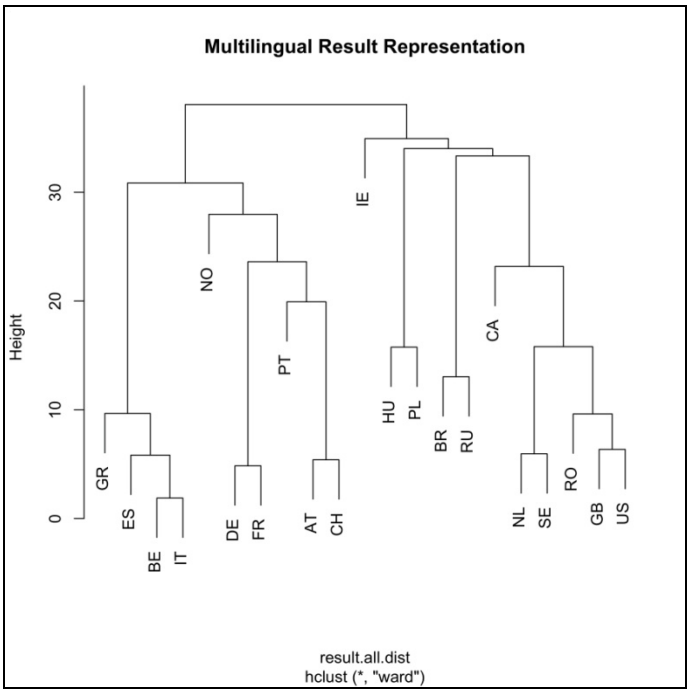

Figure 6.15 Dendogram for multilingual result representation variables

Figure 6.15 displays the country clustering with regard to the following result interaction variables:

1. *Brief Result Paging,*
2. *Full Result Paging,*
3. *Usage of Outlinks to Content Providers,*
4. *Usage of Language Facet,*
5. *Usage of Country Facet,*
6. *Selection of Native Language Facet,*
7. *Selection of Native Country Facet,*
8. *Selection of Native Language Collections,*
9. *Selection of Native Country Collections.*

At first sight, the differences between clusters are larger than it was observed for the other two digital library components. This indicates that differences between countries are stronger with regard to multilingual result interaction variables. It stands out that Norway and Ireland followed by Canada show the highest difference to other country groups. Rather similar values are computed for Belgium and Italy, France and Germany, Austria and Switzerland, Netherlands and Sweden as well as for the GB and US.

The following sections investigate similarities and differences between country groups with regard to selection of and preferences for native language and country content.

6.3.1 OCCURRENCE OF NATIVE CONTENT

Table 6.6 demonstrates the contribution of content for each country measured by content provided in the country's language(s) and by content provided from the country. Especially for countries with more than one official language, a high amount of "native" language content is presented. While Belgian institutions only provide 1.51% of the Europeana content, roughly 37% of the content is presented in at least one of Belgium's official languages. The same is true for Switzerland, Austria and Canada. Most single native language content is available for Germany and France. Less than 1% of all objects are offered for Portugal, Romania, Russia, Hungary and Greece. For Brazil, Canada and the US, no content providers contributed to Europeana, however there was content for the countries' languages available.

Country	Content per Language	Content per Country
CH	41.29%	0.47%
BE	37.56%	1.51%
CA	20.35%	----
AT	16.59%	1.85%
DE	16.59%	14.77%
FR	15.60%	15.83%
SE	10.13%	10.13%
IT	9.10%	9.10%
ES	8.38%	8.28%
NO	6.69%	6.69%
NL	5.37%	5.26%
GB	4.75%	4.71%
IE	4.75%	4.08%
US	4.75%	----
PL	4.69%	4.69%
GR	0.84%	0.96%
HU	0.52%	0.52%
RU	0.19%	0.19%
RO	0.15%	0.16%
BR	0.13%	----
PT	0.13%	0.13%

Table 6.6 Native country and language content per country

Based on the contribution displayed in table 6.6, only 7 countries have access to at least 10% native language objects. These so-called content-rich countries are Switzerland, Belgium, Canada, Austria, Germany, France and Sweden. The remaining countries are content-poor with less than 10% native language content. In a sense, Europeana is a French, German and Swedish dominated portal.

While for some countries a hugh amount of native objects are accessible, others might need to resort to non-native content to fullfill their information needs. To determine to what extent the observed countries could access native content, the three most frequent language and country facets returned for each search were extracted and analyzed with regard to native country or language occurrence. Overall, more sessions resulted in native language content (37%) than in native country content (32%). The slightly higher numbers for native language content probably results from countries with more than one official language, summarizing all as “native” sessions. Considering the relatively low percentage of native country content per country (see table 6.6), the percentage of retrieved native country objects is suprisingly high.

It can be assumed that a substantial percentage of searches retrieving native content contained queries in the assumed user’s native language. A special case are language independent queries such as “Berlin” or “Mozart” that usually retrieve objects from various providers irrrespective of the language or origin. Since documents for those queries do not necessarily appear in native language facets, no reliable assumption about the percentage of native language queries can be derived from this analysis. At least, the analysis of returned country and language facets provides an insight into the actual available native content for each session.

Not surprisingly, the content-rich countries Switzerland (69%), France (65%) and Germany (60%) received native language content more often. The least native language content was retrieved for content-poor countries like Romania (14%), Brazil (6%) and Russia (6%). The language facet "mul" as well as the highly represented German and French content was presented for queries from sessions across all countries.

Native country facets were mostly retrieved by Norwegian (78%), French (71%) and Italian users (70%), whereas Austrian (14%), Russian (7%) and Swiss (6%) search sessions rarely retrieved objects from native country providers. The high amount of native country results for Norway, France and Italy might indicate a higher usage of native language queries. A manual query analysis could prove this assumption. While Austria and Switzerland rank high for native language content, less than 2% of the available content orginate from providers from both countries.

For all countries, the facet “europe” that summarizes objects from aggegrators like “The European Library” (TEL) appeared within the search results. Additionally, objects from France and Germany are also frequently returned. Users from Austria predominantly received objects from Germany and France. According to the official languages spoken in Canada, Canadian search sessions frequently result in objects from France or the GB.

The data indicates – not surprisingly – that content-rich countries have the choice to select native content more often than content-poor countries. This might have an influence on the choice for native or non-native objects. For example, German users have a greater pool of native language objects than Russian users and therefore might select native content more often. The percentage of native content selection does not necessarily mean that German users prefer native content to Russian users. Having this in mind, the selection of country and language facets as well as assumed preferences for native language and country facets are analyzed in correlation to the available content.

6.3.2 RESULT PAGE INTERACTION

An in-depth analysis of search performance and interaction with the TEL portal identified different paging patterns across countries. The study found that users from GB, Italy, Poland and Spain are more likely to browse through result lists after posting the first query whereas users from Germany, France and the US tend to reformulate the search query instead of starting to conduct extensive result list paging (Lamm et al., 2010).

Brief and Full Result Page Interaction. For Europeana, paging behavior can either be observed within the brief or full result presentation. To which extent users page through result lists may differ according to their information needs and the number of results retrieved. A user searching for one specific object may be satisfied with the first result he assessed. A known-item search with a very specific query can also result in a very limited result list that can be scanned without any need for result list paging. In contrast, somebody searching for a broad topic like a variety of Renaissance paintings will more likely page through results and might also inspect more full result views in order to prepare a compilation of objects.

No direct interpretation of paging patterns is given here, but a trend can be observed in the specific log file sample. Figures 6.16 and 6.17 present the percentages of sessions with brief and full result paging patterns. For all countries, more full result paging sessions (34%) than brief result paging sessions (22%) were determined. This reveals that users rather choose to view single objects in detail than page through result list snippets.

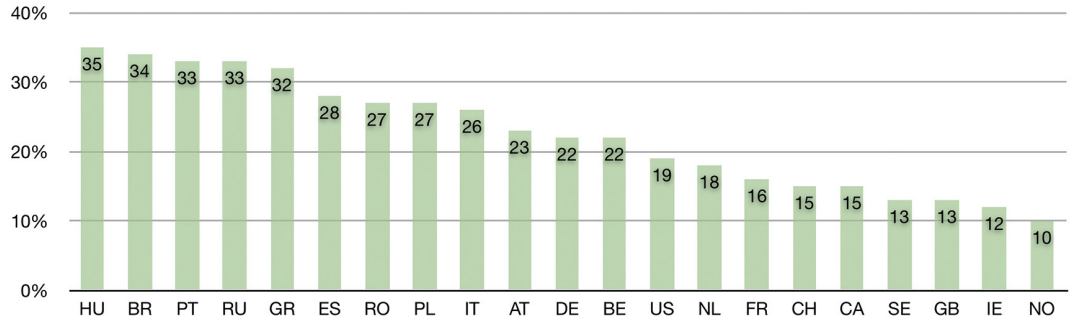

Figure 6.16 Sessions with brief result paging

Hungary (35%), Brazil (34%) and Portugal (33%) showed the most activities within brief result representations. The majority of countries that frequently paged through result lists viewed less full representations (e.g. Ireland, Brazil, Sweden, and Greece). An outstandingly contradictory pattern for brief and full result browsing is observed for Norway with the lowest brief result actions and the highest percentage of full result paging sessions. Other countries showed similar activities for both variables (Austria, Netherlands, US).

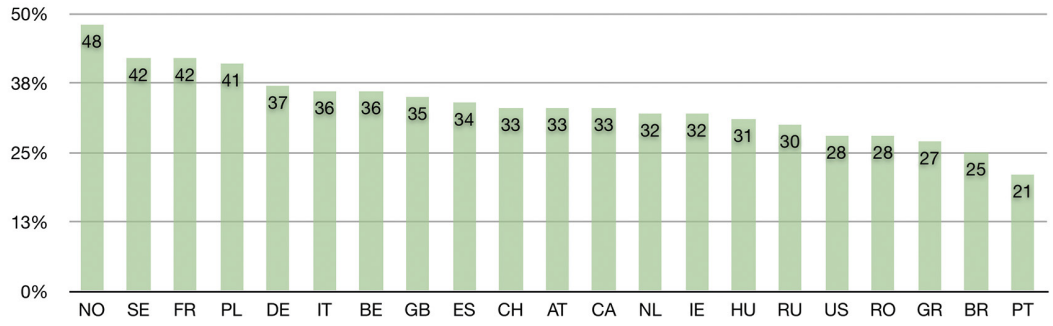

Figure 6.17 Sessions with full result paging

While Norwegian sessions contain the least result paging but are most active with regard to full views, the shortest duration of sessions was observed for this country (17 min.). In contrast, Russian sessions are the longest in duration (39 min.) with frequent result page interaction but less full views. While Norwegian users show a straightforward search path, Russian users seem to prefer the overview strategy being selective with regard to full views.

Usage of Outlinks. Having in mind that Europeana only aggregates metadata descriptions, it is necessary for users to visit the actual content provider websites in order to view the original object as well as the contextual environment of a particular object. Accordingly, it can be

assumed that visitors making use of a link to an original object view show a high interest in the retrieved result, the collection and maybe also the content provider.

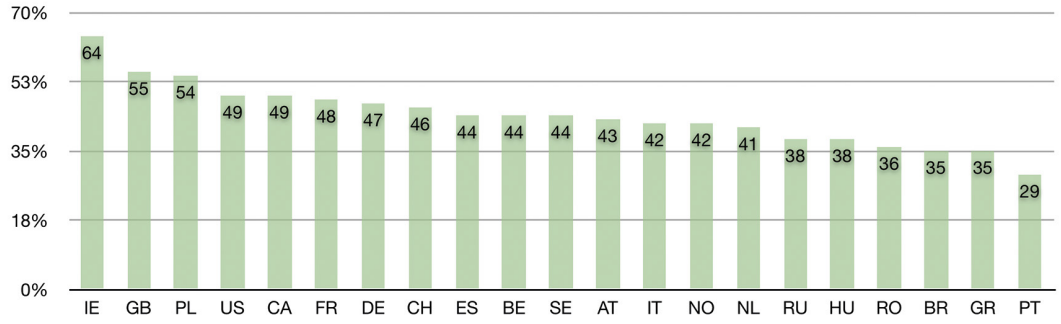

Figure 6.18 Sessions with outlink clicked

Roughly 44% of all sessions followed a link to content providers. Figure 6.18 illustrates that Portuguese sessions showed the least interest in original object views (29%). More than half of the sessions originating from the content-poor countries Ireland, GB and Poland visited content provider websites. One explanation might be that these countries expect to find more exploitable content at external websites.

6.3.3 SELECTION OF NATIVE CONTENT

While section 6.3.1 analyzed the availabilty of native content, this part of the study highlights the selection of native facets as well as objects from native collections by users in their interactions. A previous survey of European internet users has shown that users generally prefer their native language if it was available; especially users from Italy, the Czech Republic, Ireland and GB expressed their wish for native content (European Commission, 2011). Similarly, another Europeana log file study found that for some countries, stronger preferences for native collections could be determined (Clark et al., 2011).

Selection of Language and Country Facet. Result refinement through country and language facets is an important functionality for a multilingual digital library. For Europeana, both facets refer to the country of origin and language of the content provider and do not necessarily represent the language of an object or description. This classification is misleading whenever the object description is different from the language of the provider.

Only 2% of all sessions clicked on the country facet while almost 37% contained a language facet refinement (figures 6.19 and 6.20). With 3% - 4%, Greece, Hungary and Portugal made most use of the country facet. Less than 1% of all sessions from GB, Switzerland, Norway and France included a country refinement. More than half of the sessions originating from Hungary (57%), Brazil (57%) and Russia (55%) filtered results sets by language. Norway (17%) showed the least interest in refinement followed by Sweden and Ireland (21%). It stands out that content-rich countries do not show a higher usage of country or language facets. In contrast, most refinements originated from sessions from content-poor countries.

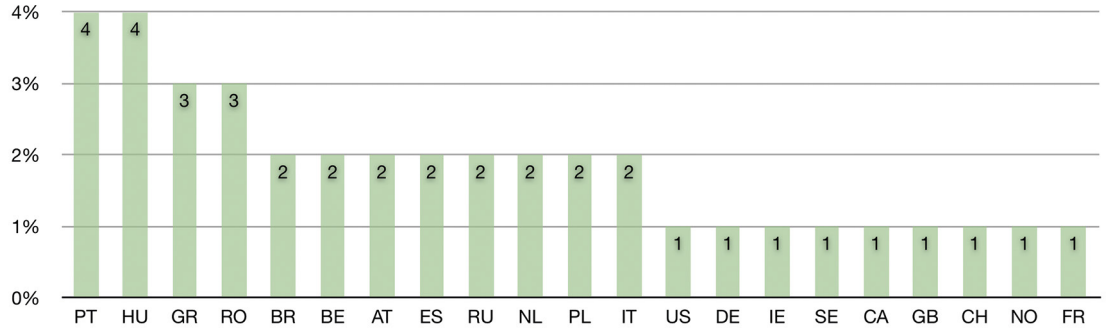

Figure 6.19 Sessions with country facet selection

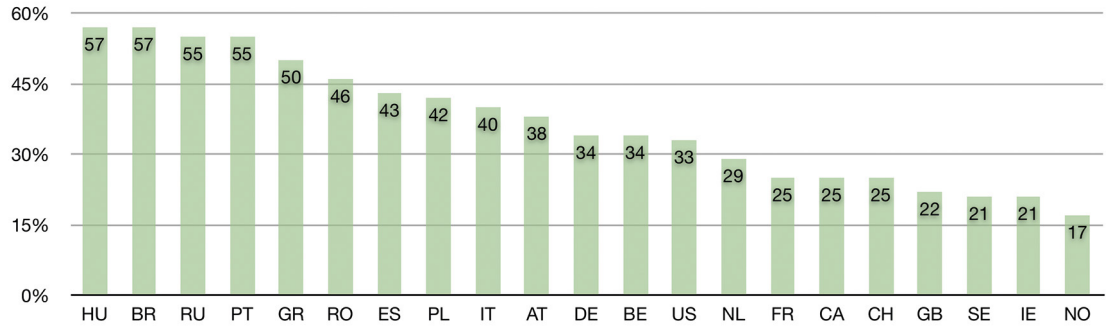

Figure 6.20 Sessions with language facet selection

With respect to the very low usage of the country facet, two conclusions can be drawn; (a) users do not show an interest in objects from their native country or (b) users do not understand the scope of the facet. The understanding and misunderstanding of facets needs to be addressed by user surveys or qualitative interviews.

For both facets, the selection of countries and languages were determined with regard to preferences for native objects. More than half of all sessions (57%) including a country facet refinement selected the appropriate native country facet (figure 6.21). While less than 1% Norwegian and French sessions made use of the country facet, they most frequently selected the facet of their native country (Norway 78%, France 71%). Again, Russia had the lowest percentage of "native" sessions (17%), followed by Switzerland (39%) and Romania (41%).

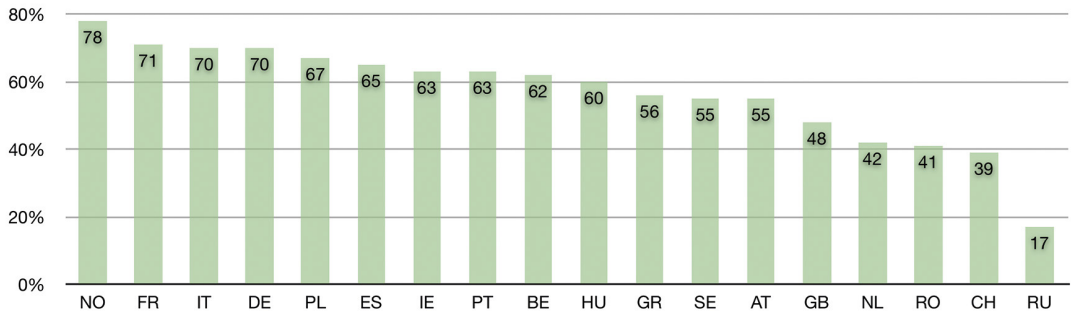

Figure 6.21 Sessions with native country facet selected

In general, a higher percentage of sessions (62%) refined searches according to their native language(s). Content-rich countries select native language facets more often than content-poor countries. The French and German speaking countries Switzerland (86%), France (86%), Germany (84%) Belgium (84%) and Austria (82%) predominantly selected their native language facets. Romania (31%) and Russia (16%) together with Sweden (41%) tail the field of "native" sessions (figure 6.22). For content-poor countries, the English facet was used more frequently. This serves as an indication that especially content-poor countries use English as secondary language to find results. While less than 50% of the American sessions choose their native language facet, a frequent selection of the French facet was observed.

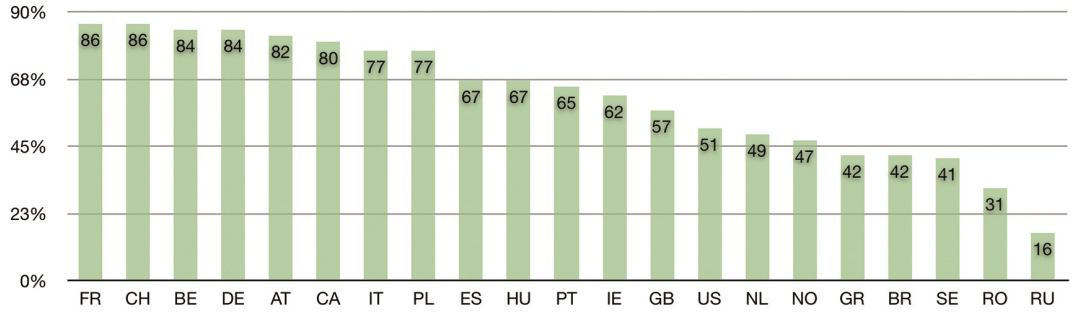

Figure 6.22 Sessions with native language facet selected

Language and Country of Collections viewed. So far, it has been investigated to what extent users are presented with native content and if they tend to filter results by (native) language and / or country facets. Additionally, the selection of objects can give an insight into the relationship between the usage of native and non-native content. As mentioned above, language and country facets refer to the collection's origin. For this study, the language of an object is defined by the language of its collection. Whenever a user clicks on an object from native language or country providers it is counted as a selection of native content.

According to the data at hand, users selected native country objects in 46% of all sessions and native language objects in 43% of the sessions.

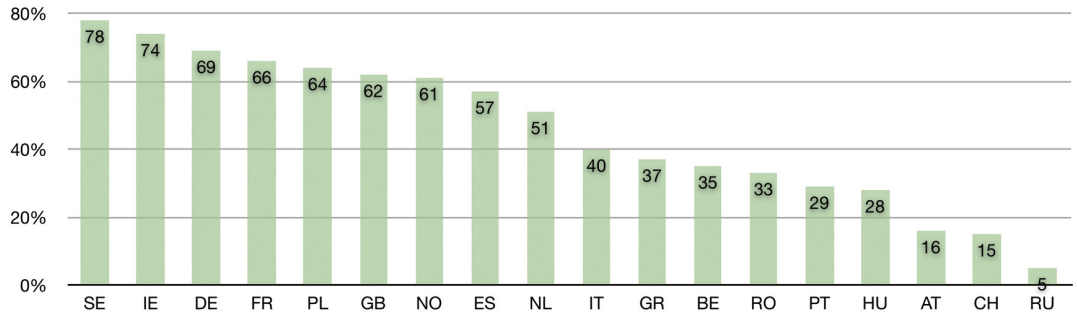

Figure 6.23 Sessions with native country collections selected

In contrast to the low country facet usage, at the actual object level users show more interest in native country objects. The content-rich countries are once more at the top, selecting native objects more often than content-poor countries. The vast majority of Swedish (78%), Irish (74%) and German (69%) sessions looked at objects provided by their countries (figure 6.23). Alligned with the available native country content, only a few sessions from Austria (16%), Switzerland (15%) and Russia (5%) contained a selection of native objects.

Similar to the country preferences, Swedish (77%) and German (67%) sessions tend to select native language objects more often (figure 6.24).

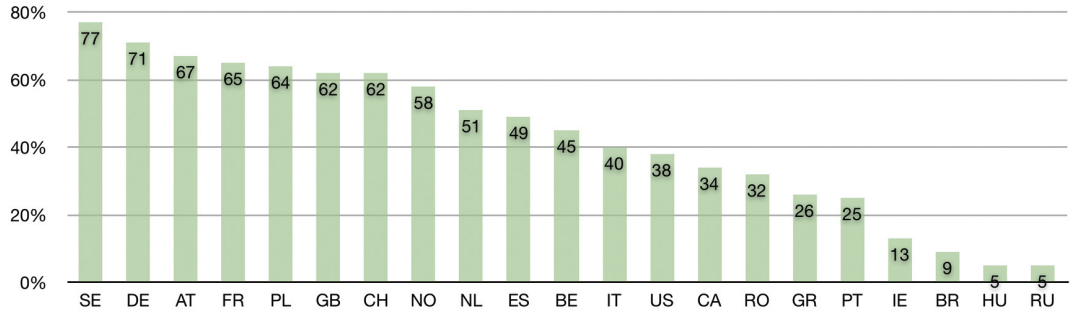

Figure 6.24 Sessions with native language collections selected

Especially for the smaller countries with more than one official language, the origin of an object or collection is less important than the object's language. Looking at the example of Austrian sessions, it is obvious that those users very rarely choose objects from native country providers, but show a strong interest in native language objects (67%). Less than 10% native language object views are counted for Brazil (9%), Hungary (5%) and Russia (5%). For content-poor countries, a tendency for German and English collections as well as for multilingual collections from aggregators like TEL is observed.

6.3.4 COMPARISON OF COUNTRY PAIRS

With regard to multilingual result representation interactions, one may reasonably expect that content-rich countries differ most from content-poor countries. As mentioned before, the available native content is expected to play an important role when accessing cultural heritage objects. Taking all variables into account, no significant separation of content-rich and content-poor countries can be confirmed, but a trend for content-rich countries to select native language facets and content is observed. For content-poor countries, a small interest for native country facets comes with a higher interest in native country objects.

Country differences are determined using the Marascuilo procedure reporting the deviation between the absolute (AD) and critical difference (CD). Results for all country comparisons can be found in Appendix E.

Statistically, the difference between 201 out of 210 country pairs was significant for the selection of objects from native language collections and for 147 from 153 country pairs with regard to the selection of objects from native country collections. For both variables, the

differences between Germany as a content-rich country frequently selecting native content and the content-poor country Russia with rare native object selections were most significant (Native Language Collection, AC/DC= 43.17), (Native Country Collection, AD/CD= 44.48). Similarly, sessions vary with regard to brief result page interaction (183 out of 210). The most contradictory behavior is shown between Norwegian sessions with more than 95% and Spanish sessions with less than 30% paging sessions (AD/CD=12). In comparison to brief result page interactions, sessions are more similar with regard to full object views with 143 signficant differences out of 210 country pairs. Again, users from the content-rich country France more often click on an object than users from the content-poor Portugal (AD/CD=7.44). Irrespective of the available content, sessions differ in 178 out of 210 cases for the usage of outlinks. More than half of all sessions from Poland and less than 30% of Portuguese sessions used an outlink to view objects at the content providers website.

Less variations are present with regard to the usage of the language (153 out of 210) and country facet (130 out of 210). For the selection of the native country facet only 54 out of 153 and for the language facet 120 out of 210 country pairs show significant diverse characteristics. Differences appear for outliers like France with frequent selections and very few selections of the native country facet by Russia (AD/CD=3.55).

6.4 RANKING OF VARIABLES

The previous analysis put the focus on differences between country groups providing an answer to the question whether sessions characteristics from different countries vary with regard to selected variables. So far, all variables have been equally treated assuming that all of them have the same impact on interactions. In this section, the value of each investigated variable for country and language specific logging is evaluated. As a result, a list of strong and weak variables for country and language specific interactions and preferences evolves, answering research question three:

RQ3: Which variables gathered by log files uncover significant country and language specific differences in user interactions?

In order to determine strong and weak variables, a cluster analysis taking all variables from each of the three digital library components into account can serve as a basis for the classification. Strong variables are those that have an impact on the cluster analysis. Weak variables show similar values for each cluster and therefore only have a minor influence on the analysis.

The *k*-means algorithm estimating mean values for a set of predefined groups or clusters was conducted identifying high and low impact variables for country and language level differences between observations. The algorithm groups observations in iterative steps to their nearest mean or median cluster group. For the *k*-means clustering algorithm, the number of cluster *(k)* needs to be determined. As discussed in chapter 5, the elbow method was applied comparing the sum of squared distances (SSD) for sequenced cluster solutions (Ketchen and Shook, 1996). Based on the elbow criterion as well as the results from the Ward's hierarchical clustering, the number of clusters was determined in iterative steps optimizing *k* for the three digital library components.

Selected variables can be classified as high or low impact variables based on the cluster variances. In other words, variables with at least 30% variance from cluster medians represent differences between country characteristics. Variables with less than 30% variance indicate similar patterns of behavior and are therefore not as important identifying country and language specific variances for the development of effective multilingual digital libraries.

Digital Library Component	Number of Clusters
Multilingual interface	3
Multilingual search and browsing	4
Multilingual result representation	4

Table 6.7 Number of clusters for digital library component variables

For the three components, 3 to 4 clusters were determined as most the purposeful division (table 6.7). Based on the predefined number of clusters, the cluster analysis is conducted for each component in order to determine which variables have a high impact on country and language level differences. For each variable, cluster medians are computed. Median values are compared since they are more robust with regard to outliers. Values for each cluster are based individual country values belonging to this particular cluster. High impact is defined by the variation of median cluster values per variable. Cluster values that are deviating for more than 30% from all medians for one variable are underlined.

Table 6.8 contains the median values from all country groups for the four multilingual interface variables. The first column contains the number of clusters with the second column presenting the number of countries for each cluster. In this case, the majority of countries (12) are summarized in cluster one. Cluster two contains five countries and cluster three four countries. The remaining columns present the median values for each cluster per variable.

For native language browser version usage, all three values are similar. This indicates that all countries behave similarly and choose their native language browser version. The second cluster

of the native Google version usage deviates most from the two remaining cluster values. For the third cluster, a strong deviation is obvious within the language change variable caused by a median value of 0.98 compared to 0.12 and 0.23. Similarly, countries within cluster three vary from the other two cluster groups with regard to the selection of the native interface language. Consequently, the variables Google language, interface language change and selections of native interface language indicate a strong language impact.

Multilingual Interface					
Cluster	**Number**	**Browser Language**	**Language of Google Referrer**	**Interface Language Change**	**Interface Language**
1	12	0.92	**0.15**	**0.12**	0.69
2	5	0.85	**0.32**	**0.23**	**0.47**
3	4	0.92	**0.07**	**0.98**	**0.96**

Table 6.8 Results for multilingual interface variables cluster analysis. Cluster values that are deviating for more than 30% from all medians for one variable are underlined.

For the multilingual search and browsing cluster analysis, a relatively homogenous distribution is observed for most variables (table 6.9). The elbow test suggests distribution of the observed countries among four clusters. Differences, albeit small, between the country groups were calculated for the variables unique queries, login (LG) and session duration. The results for session duration and queries each consist of two similar clusters. The median number of queries is similar for clusters 1 and 4 (1.9) as well as for clusters 2 and 3 (1.6). Cluster 1 of the session duration has a higher value (26.65) than the remaining clusters groups but without significant differences. With regard to the user profile usage, cluster 4 contains twice the number of sessions (0.12) with a log in. Hence, the two countries within this cluster more often log in the user profile than the other observed countries and therefore differ most to the remaining country values for the user profile usage.

Multilingual Search and Browsing								
Cluster	**Countries**	**External Access Point**	**Bounce Rate**	**Login**	**Duration**	**Unique Queries**	**Search Sessions**	**Browsing Sessions**
1	6	0.88	0.16	0.05	26.65	1.93	0.80	0.02
2	8	0.89	0.16	0.06	23.60	1.65	0.79	0.03
3	5	0.88	0.17	0.05	20.92	1.66	0.83	0.02
4	2	0.86	0.12	**0.12**	20.22	1.95	0.74	0.04

Table 6.9 Results for multilingual search and browsing variables cluster analysis. Clusters values that are deviating for more than 30% from all medians for one variable are underlined.

In total, nine variables are included in the multilingual result representation component analysis (table 6.10). The countries are divided into four clusters with the majority of the countries in the first two clusters. Besides the two variables full result paging and usage of outlinks, all remaining variables show significant variances between the cluster groups. For the usage of outlinks, only the last cluster containing one country shows a difference of more than 30% in comparison to the remaining cluster values.

Countries within cluster 3 and 4 browse more often through result lists. With regard to the usage of language and country facets, cluster 3 is characterized by a higher usage of both facets. Cluster 2 mostly selects native language facets in contrast to the lower usage of native country facets and collections. Country groups from cluster 1 show most interest in native language collections.

Multilingual Result Representation										
Cluster	**Countries**	**Brief Result Paging**	**Full Result Paging**	**Language Facet**	**Native Language Facet**	**Country Facet**	**Native Country Facet**	**Native Language Collections**	**Native Country Collections**	**Outlink**
1	10	0.22	0.35	0.36	0.79	**0.02**	0.64	**0.53**	0.38	0.43
2	8	0.18	0.31	0.31	**0.45**	0.01	**0.29**	0.36	**0.19**	0.42
3	2	**0.31**	0.36	**0.49**	0.72	**0.03**	0.63	0.34	0.46	0.46
4	1	**0.12**	0.32	**0.21**	0.62	0.01	0.63	**0.13**	**0.74**	**0.64**

Table 6.10 Results for multilingual result representation variables cluster analysis. Cluster values thar are deviating for more than 30% from all medians for one variable are underlined.

Summarizing, substantial differences were calculated for the Europeana interface language change and the selection of native interface languages as well as for the selection and preference of native content. Table 6.11 compiles an overview of high and low impact variables in descending order of impact according to the number of clusters with large differences.

Digital Library Component	High Impact Variables	Low Impact Variables
Multilingual Interface		
	Europeana Interface Language Change	Browser Locale Language
	Europeana Interface Languages	
	Language of External Google Referrer	
Multilingual Search and Browsing		
	Login	External Access Point
		Bounce Rate
		Search Sessions
		Browsing Sessions
		Duration of Sessions
		Unique Queries per Session
Multilingual Result Representation		
	Selection of Native Country Facets	Full Result Paging
	Selection of Native Language Facets	Outlinks to Content Providers
	Country of Collections	
	Language of Collections	
	Usage of Language Facet	
	Usage of Country Facet	
	Brief Result Paging	

Table 6.11 High and low impact variables for each digital library component

From the 20 investigated variables, eleven have a high impact and nine variables have a low impact on the evaluation of country characteristics in multilingual digital libraries. Not surprisingly, variables related to the interface language usage and selections of objects are the strongest indicators for country and language level differences.

The strongest variations between country groups are detected for the Europeana interface language change followed by the selection of native country and language facets or content.

Country and language specific studies within multilingual digital libraries need to consider high impact values to identify similarities and differences between country, language or even cultural groups. Depending on the research question and system under investigation, high impact variables can differ or be extended by low impact or additional variables not considered in this study.

6.5 SUMMARY

The null hypothesis that all countries show the same usage patterns was rejected. For all variables and in particular for interface and content related interactions, significant differences exist between the countries.

It was determined that users from most countries make use of native language interface versions in daily life. A clear preference for native language interfaces regarding browser locale and the Google search engine was observed. In contrast, a different behavior was detected for the Europeana portal interface language. Only a minority of users switched the interface language to their assumed native language – maybe because of the infrequence of use it does not appear necessary or maybe because switching the interface language in the portal is too troublesome. From the analysis it is not entirely clear whether Europeana meets the user needs for native language interfaces or should rather draw upon the experiences with general search engines like Google.

Seven variables were investigated focusing on multilingual search and browsing patterns. All countries tend to access Europeana via external access points and rather use the search box than browsing features. Very little interest is observed for the user profile. While Norwegian users show a straightforward search path, Russian users seem to prefer the overview strategy being selective with regard to full views.

The qualitative analysis of two sample query sets from Germany and France has shown that the majority of searches are looking for named entities that are often language independent. The content of queries seems to be country specific – especially in a cultural heritage portal like Europeana. The duration of sessions and the number of unique queries per session also varies slightly between the country groups – possibly dependent on the nature of the content (i.e. language) available for the user.

According to the available content within Europeana, countries are considered either content-rich or content-poor. For this study, countries are considered to be content-poor if they feature

less than 10% native language content while content-rich countries hold more than 10% native content within Europeana.

The data indicates – not surprisingly – that content-rich countries have the choice to select native content more often than content-poor countries. This might have an influence on the choice for native or non-native objects. However, taking all variables into account, no significant separation of content-rich and content-poor countries can be confirmed, but a trend for content-rich countries to select native language facets and content is observed. For content-poor countries, a small interest for native country facets comes with a higher interest in native country objects. For content-poor countries, a tendency for German or English collections and for multilingual collections from aggregators like TEL is observed as well as more frequent visits to content provider websites. One explanation might be that these countries expect to find more exploitable content in foreign languages or at external websites.

The degree of influence of single variables on country and language differences was determined using a cluster analysis approach. From the 20 variables, 11 are classified as high impact indicators. The strongest distinctive features are the usage of the Europeana interface language (change) as well as the usage and selection of (native) language facets and content.

While the previous analysis identfied differences between countries based on single variables, chapter 7 focuses on country characteristic describing profiles. Single country and country group profiles are visualized and compared with regard to attributes based on all investigated variables.

7. COUNTRY PROFILING

The analysis of interactions from different countries provides an insight into country and language level differences. At the same time, countries can be characterized with regard to the observed interactions and investigated variables. This is the goal of this chapter. Similar to Hofstede's five dimensions model (Hofstede, 2010), countries are explored using the 20 variables representing multilingual digital library usage dimensions.

For the graphical illustration of country profiles, radar diagrams were chosen. Each spoke represents one variable. Exemplary comparisons are drawn between all countries, between two individual countries, an individual country to an averaged country profile, content-rich and content-poor countries and English and non-English countries (averages over individual profiles). The visualization of results is a helpful mechanism to categorize, group and compare country characteristics with regard to the observed variables and inform future directions in multilingual digital library development.

7.1 COUNTRY PROFILES

In chapter 6, interactions were compared with regard to country level differences based on single variables. Based on the analysis in chapter 6, visualizations can either focus on differences between the complete sample set, on selected country groups or on single countries.

As an example, figure 7.1 displays to what extent countries make use of the native language Google version. Green countries show the highest usage while red countries only rarely make use of native language versions. At first sight, Canada, France and Germany show the highest usage and Russia the lowest selection and usage.

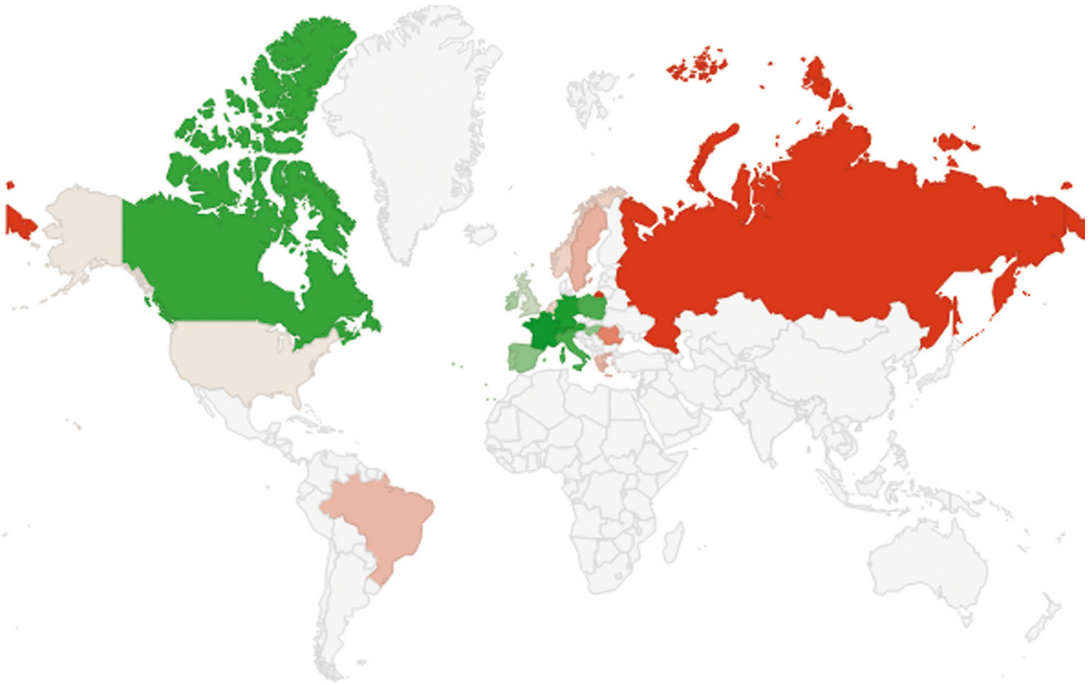

Figure 7.1 Visualization of a single variable (Google Referrer) for all countries

While this map visualization provides a good overview of single variables, it does not allow describing single countries and their attributes. From figure 7.1, Russia stands out as lowest native language interface usage country with regard to the Google search engine. But does Russia in general accept default or non-native interface settings and content? In order to answer these questions, a country profile for Russia could be consulted.

A consolidated view of all variables leads to country profiles that can be used to describe interactions within a particular system under investigation. Focusing on a single country, profiles emerge representing country characteristics derived from the investigated variables. A table containing values for each country as well as the visual representations can be found in Appendix A.

The visualization through radar charts simplifies the identification of similarities between several observations as well as dissimilarities and outliers. Since radar charts require values on an equal scale, the two variables duration of session and unique queries per sessions are not included in the visual representation but taken into account for the description of each country or comparison of country groups.

Figure 7.2 presents Russian session characteristics for the 20 variables discussed in chapter 6. Clockwise, variables from the multilingual interface, multilingual search and browsing and multilingual result representation components are displayed.

Comparable to Hofstede's model that describes cultural tendencies with regard to 5 dimensions, the country profiles identify country and language specific usage tendencies with regard to 20 dimensions. The profile for Russia indicates an active country with a long session duration, a high number of queries and extensive result page interactions. More than 50% of all Russian sessions make use of the language facet but only 16% selected Russian as their preferred result language. In contrast, only 2% of all sessions used the country facet from which 17% selected the Russian facet. Effectively, only a small percentage of sessions viewed native country or language collections. An explanation for this might be the low number of available content from Russia or in the Russian language.

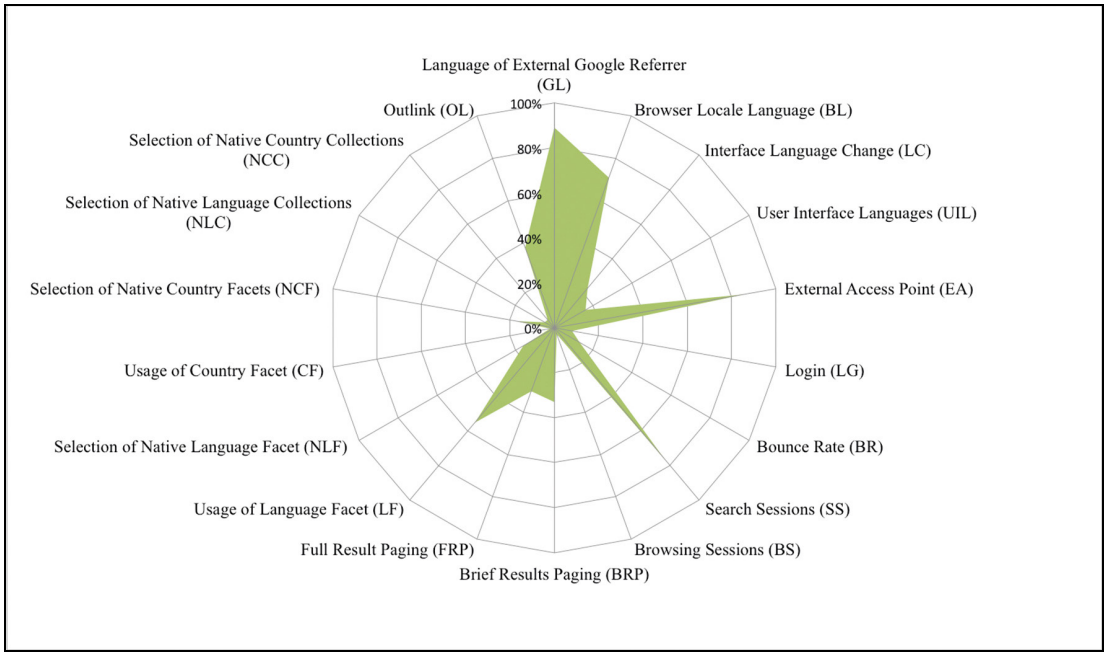

Figure 7.2 Russia country profile

Comparing the Russian country profile to a country with different interaction patterns uncovers the range of behavior. Using the example of Poland as another content-poor country, a contradictory trend with regard to result interactions is observed (figure 7.3).

In comparison to Russia, Polish sessions refine results less frequently by language but more often select their native language as well as their native country facet. Similarly, Polish sessions show a strong selection of native language and country facets in contrast to Russian sessions.

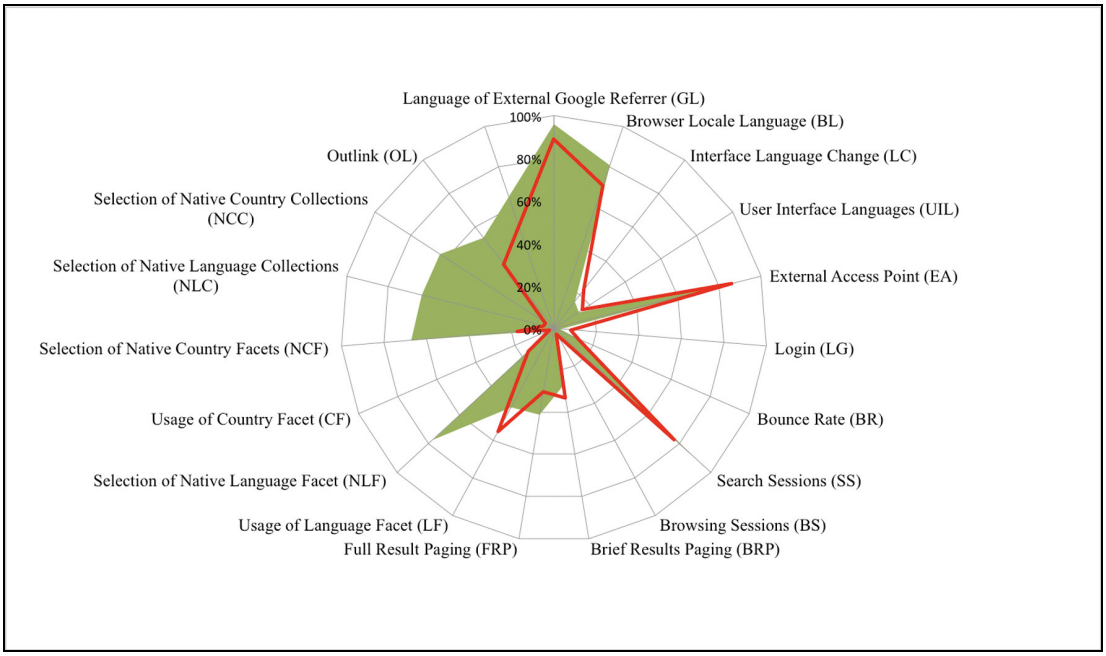

Figure 7.3 Poland country profile (green) and Russia country profile (red)

Country profiles can either represent single countries like shown above or groups of countries for meta analysis and comparisons. Country characteristics can be summarized under different aspects such as languages, available content or location. As an example, the following section compares:

- An individual country (here: France) to the median country profile,
- Content-rich and content-poor countries (median to individual profiles),
- English and non-English countries (median to individual profiles).

7.2 MEDIAN COUNTRY PROFILE COMPARISON

To evaluate whether a country has specific characteristics, a median country profile was produced, which is represented in figure 7.4. This profile represents an average over all individual country profiles in order to provide a comparison to a mean. The diagram shows that averaged over all countries, a high preference for native language Google (GL) and browser versions (BL) and in comparison a very low usage of the Europeana interface language change (LC) or sessions with the native interface language (UIL) selected can be detected. The majority of sessions were directed from external access points (EA). Most users conducted at least one

search (SS), mainly using the search box rather than browsing entries (BS). On average, sessions last 24 min. (D) and contain 1.76 unique queries (Q) (not represented in the diagram).

Search results are mainly filtered via the native language facet (NLF). The small number of sessions containing a country facet refinement (CF) predominantly selected their native country (NCF). The interest in full object views (FRP) tends to be higher than in result list paging (BRP). Less than half of all sessions clicked on the content provider link to view the original object (OL).

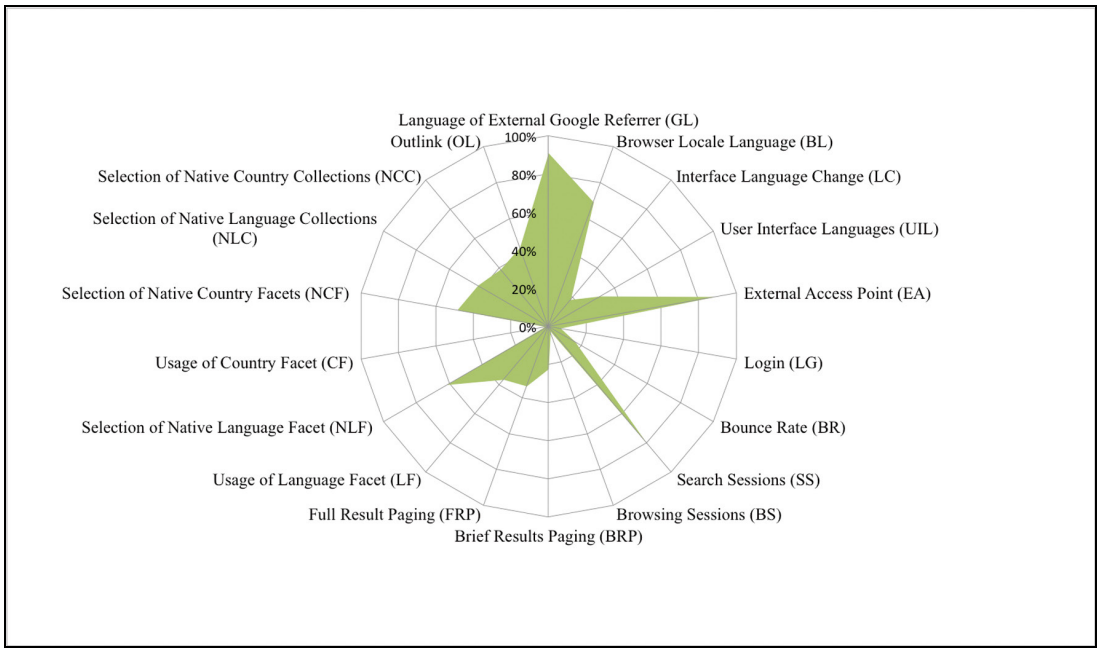

Figure 7.4 Median country profile (median of all individual profiles)

Figure 7.5 compares the median country values with characteristics from French sessions (red) as the country with most accesses within the dataset. French sessions are above the average for the usage of native interface versions for Google (GL) and native language browser versions (BL). Opposite are a lower usage of the Europeana interface language change (LC) and fewer sessions with the native interface language (UIL). French sessions do not considerably differ from the mean values with regard external accesses points (EA), the usage of a user profile (LG), bounce rates (BR) and browsing patterns (BS). With 90% search sessions, France is 10% above the median value for all countries. With regard to paging behavior, French users show more interest in full result views (FRP) and fewer result page interactions (BRP).

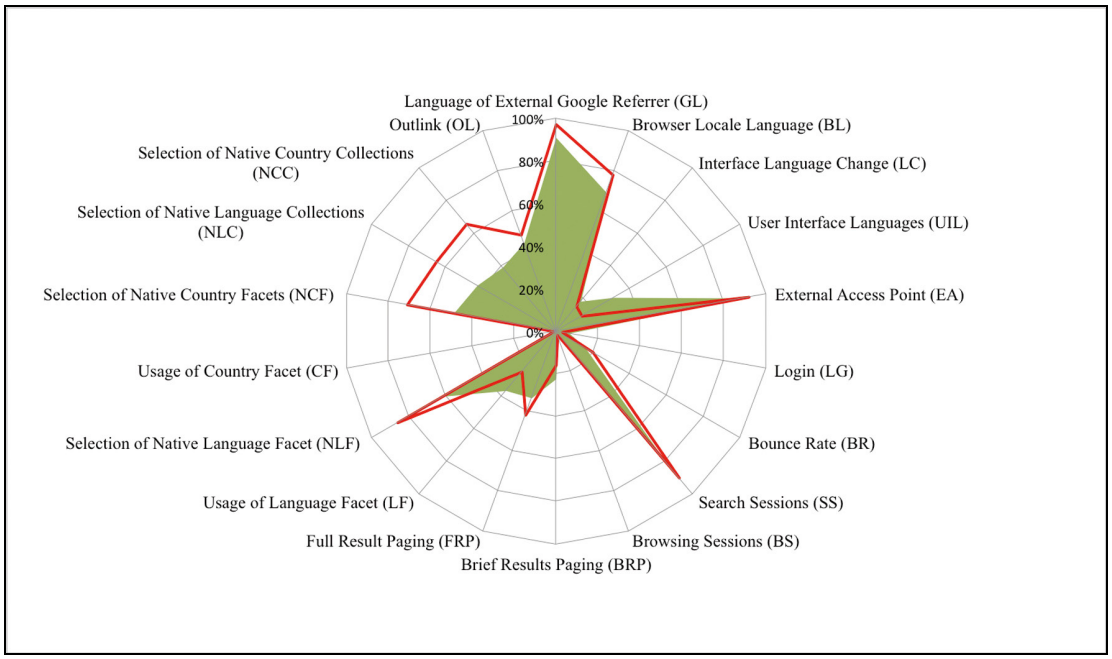

Figure 7.5 Median country values (green) and French sessions (red)

Most differences are determined for result representation variables. While French sessions make use of facets less frequently, French users select their native language (NLF) and country facet (NCF) more often than observed for median country values. Above average, France users select native country (NCC) and language objects (NLC) as well as outlinks to content providers (OL). A reason for those findings might be that France is a content-rich country with 16% native language content available in Europeana.

The examination of single countries compared to median usage trends provides an insight into individual country or language patterns that can be leveraged for the improvement of system design as well as content aggregation and representation.

7.3 CONTENT-RICH VERSUS CONTENT-POOR COUNTRIES

In chapter 5, a distinction between content-rich and content-poor countries was made. In particular with regard to multilingual result representation options, the comparison of content-rich and content-poor country characteristics provides an insight into major differences indicating possible improvements and support functionalities for content-poor countries. Figure 7.6 presents profiles for content-rich countries displayed in red and content-poor countries in green.

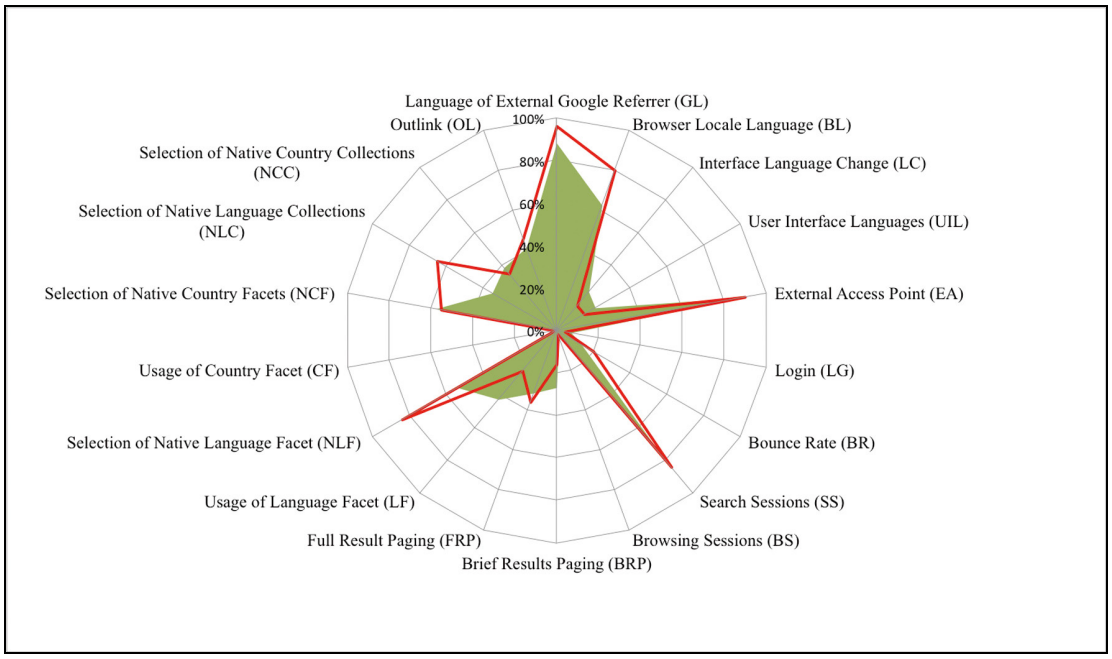

Figure 7.6 Content-rich (red) and content-poor countries (green)

With 96% Google native language (GL) and 79% native browser version (BL) accesses, content-rich countries show a higher preference for native language applications compared to content-poor countries. Interestingly, the opposite is true for the Europeana interface language change (LC) and the selection of the native interface (UIL). Content-poor countries more often change the interface language to their native version. This might be a result of content-poor countries being used to higher efforts to access websites in their native language. For search (SS) and browsing (BS) variables, only very small variances exist. The comparison of session duration (D) and unique queries (Q) indicates that content-poor countries need slightly more time (24 min.) on average and type more queries (1.8) than content-rich countries (22.5 min; 1.65 queries) to find (relevant) results. This confirms the assumption that content-rich countries have an advantage in terms of efficient searches over content-poor countries that need more effort to find relevant results.

With respect to paging behavior, contradictory patterns are observed with more result page interaction (BRP) from content-poor countries and more full views (FRP) for content-rich countries. Based on this pattern, it can be assumed that content-rich countries more often find results directly while content-poor countries browse through result pages to find relevant objects. Content-poor countries tend to refine search results via language facets (LF) but a higher selection of the native language facet (NLF) is determined for content-rich countries. Almost twice as many sessions from content-rich countries access native language content

(NLC) compared to content-poor sessions. Concerning the usage of country facets (CF) and selection of native country facets (NCF) and objects (NCC) as well as the usage of an outlink (OL), no variation between the groups can be identified.

The comparison has clearly shown that the availability of native language content influences interaction patterns. While content-poor countries showed an active interest in language refinement, the missing content leads to lower interaction with objects.

7.4 ENGLISH VERSUS NON-ENGLISH COUNTRIES

While English is the predominantly used language for website content in general, Europeana is not an English dominated portal. Therefore, Europeana is suitable for the comparison of English and non-English usage data (figure 7.7). Only for the interface language, English is used as default setting. Not surprisingly, almost 100% of sessions from English speaking countries were conducted with the English interface language (UIL) (green). Non-English countries still rarely change the interface language to their native version (red). A difference is also present for the usage of native browser versions (BL), with 97% native browser accesses for English countries compared to 63% native browser accesses from non-English countries.

While slightly more search sessions are coming from English speaking countries, they are shorter in duration (D) (19 min.) and contain only 1.5 unique queries (Q). In comparison, non-English sessions last 24 min. and contain 1.8 queries. On average, non-English sessions are five minutes longer than English sessions which are explainable with a higher result page interaction. English speaking countries might have fewer users with foreign language skills and therefore fewer interactions with non-native content.

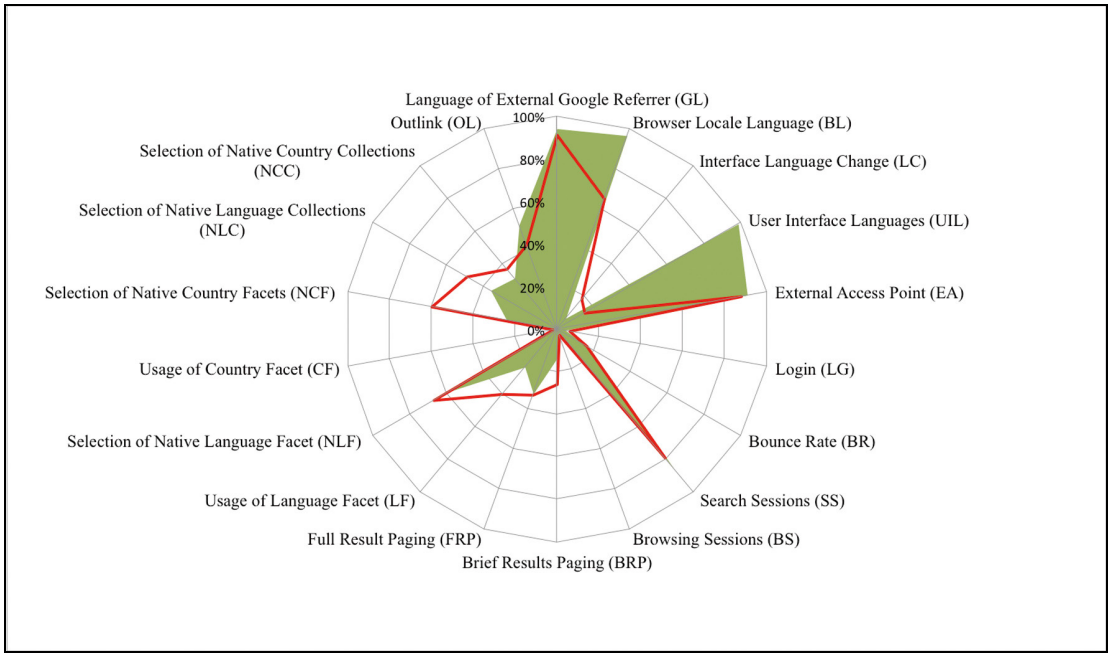

Figure 7.7 English (green) and non-English countries (red)

Non-English countries tend to page through result lists (BRP) more often but do not differ from English sessions with regard to full views (FRP). With respect to the selection of facets, non-English sessions more often refine results via the language facet (LF) without more native language facet selections (NLF). In comparison to only 28% native country facet selection (NCF) for English sessions, almost twice as many non-English sessions selected their native country facet. In view of selecting objects, non-English sessions show a slightly stronger preference for native language (NLC) and country objects (NCC). English sessions often view objects at the content provider's website (OL). An explanation for this could be the fact that the English-speaking countries count among content-poor countries with fewer native-language content within Europeana.

7.5 SUMMARY

The identification of country differences as well as country characteristics is the basis for effective multilingual digital library development. Through the development and comparison of profiles, differences between single countries and country groups can be highlighted.

Knowing about country or language specific roles and requirements, purposeful decisions can be made regarding interface design, search and browsing functionalities as well as content aggregation.

The presented profiles act like “personas” representing countries or country groups. Country personas can be used for the evaluation of existing digital libraries as well as for user requirements analysis. The comparison of country profiles from other multilingual digital libraries determines to what extent characteristics are cross-system or system specific. Based on the country data and visualization options, several other comparisons and presentations are possible.

8. CONCLUSION

The investigation of interactions demonstrated that country and language level differences exist and need to be taken into account when designing effective multilingual digital libraries serving international users.

This chapter summarizes the main outcomes of this research and highlights additional findings and recommendations for MLIA within the context of previous research. The dissertation concludes with an outlook on future and complementary work in the field of user studies in multilingual digital libraries. The focus lies on purposeful correlations, the impact of the interface language change and native content on user interactions.

8.1 RECOMMENDATIONS FOR MULTILINGUAL DIGITAL LIBRARIES

The observed interactions and patterns partly confirm previous outcomes and recommendations as discussed in chapter 2. Additional results and recommendations are provided especially for multilingual interface usage and result representation.

For multilingual interface usage, the two predominantly used localization options were examined: (1) automatic interface language (change) based on location (i.e. Google) and (2) default English interface with user-triggered language change options (i.e. Browser and Europeana interface language). In line with previous findings, a low usage of the Europeana interface language change is observed. In contrast, the examination of Google language parameters has shown that the majority of sessions originated from their native language Google version. Similarly, users tend to use native language browser versions rather than English versions. Based on these findings, two assumptions could be made. First, users seem to accept the Europeana default English language settings as well as automatic changes to their native language interface. Second, users prefer and select their native language versions for frequent usage. Previous studies of Europeana visits have shown that most users are first time visitors while only a few are returning visitors. Therefore it is plausible that for most users an interface language change for the Europeana portal is not worth it due to the short-time usage.

More research is needed to examine the preferences for native language interfaces in different settings. Additionally, the intention, assumptions and maybe even misunderstandings related to

the interface language change need to be investigated through additional qualitative user studies. From previous log file studies, the question arises whether users expect a connection between the interface and query language. Examples as well as an outlook for future work on this are presented in the next section.

All countries show similar patterns for access points with a majority of sessions directed from external access points. In particular, many users are guided by search engine result pages referring to the Europeana portal. Since the analysis of Google referrers has shown that the majority of users are directed from their native language version, the appropriate Europeana version could rank first in the result list (increasing the native language interface use in Europeana). Currently however, the default English version ranks before the language specific version.

None of the observed countries shows a high usage of the user profile "MyEuropeana". The currently available low level personalization options do not play an active role within search sessions. The provision of a user profile containing more (collaborative) features can attract users to engage with the portal and its content. This is especially important since user background information as well as user-generated content can be leveraged for country or language specific personalization. For Europeana, the storage of the user's native and preferred language can improve search experience by personalized search suggestions as well as result representations.

On average, the analyzed sessions are longer in duration (24 min.) and contain more queries (1.8) than results from previous studies of TEL and Europeana logs indicate (Angelaki, 2007; Clark et al., 2011). The average duration and number of queries varies between the country groups. A correlation between a relatively long session length and number of unique queries was observed for some countries like Russia and Poland. In contrast, the Netherlands shows the second longest session length (30.6 min.) but only 1.7 unique queries. Similarly, Hungarian sessions are rather short in duration (21 min.) but contain 2 unique queries on average. Systems should be flexible with regard to different search and browsing procedures.

Interestingly, Norwegian sessions are the shortest in duration (17 min.) and contain the least result paging but are most active with regard to full views. In contrast, Russian sessions are the longest in duration (39 min.) with frequent result page interaction but less full views. Since both countries belong to the group of content-poor countries, the difference cannot be explained by the available content but might be a result of different search strategies. While Norwegian users show a straight forward search path, Russian users seem to prefer an overview strategy being selective with regard to full views.

In line with previous studies, the qualitative analysis of queries from Germany and France showed that the majority of searches is for named entities whose language is often ambiguous. This is a particular challenge for query language detection. Reliable query language detection is a crucial step for the provision of cross-lingual search. While both countries displayed a high coverage of named entities in queries, differences were observed with regard to the query content.

When interacting with content, content-rich countries more often select native language facets and native language objects. Considering a minimum of native language content per supported language or English translation of the largest collections could therefore be helpful for content-poor countries. However, the usage and selection of (native) country facets does not play an important role. Consequently, one might draw the conclusion that users are interested in native language content but not in the origin of objects. Table 8.1 summarizes overall trends and patterns derived from this study and suggests possible system design consequences for each digital library component.

Component	Outcomes	Recommendation	Section
Multilingual Interface			
	Usage of native interface versions for frequently used systems.	Draw upon experiences with frequently used systems, possible direction to appropriate Europeana version based on referrer or browser language.	6.1.1
	Users rarely switch the Europeana interface language.	Provide English interface as default or automatic interface language change.	6.1.2
	Drop-down interface language change is the most common interface language option used.	The user should be able to change the default or automatic selected interface language during a session.	6.1.2
Multilingual Search and Browsing			

	The majority of sessions are directed from external links (most often from search engines).	Use language parameters from external links for interface language change.	6.2.1
	Sessions from different countries show different session lengths and number of queries per session.	The system should be flexible with regard to different search strategies.	6.2.2
	Searches are mainly for named entities.	Provide named entity recognition before query translation.	6.2.3
	Queries are often language independent.	Provide language identification of queries through user selection.	6.2.3
	Users from different countries have different, partly national or regional information needs.	Provide country specific search suggestions.	6.2.3
Multilingual Result Representation			
	Brief result and full result paging differs between country groups.	Support different result inspection strategies.	6.3.2
	Users frequently refine results by language.	Provide language identification of all objects and refinement via advanced interface and facets.	6.3.3
	Users rarely refine results by country of origin.	Focus on the object language.	6.3.3
	Content-rich countries show higher preferences for native content than content-poor countries.	Consider higher ranking of native content. Provide balanced	6.3.3

		native language content for each supported language.	
	Content-poor countries select more English content.	Provide English metadata translations, support content-poor countries with translation tools.	6.3.3
Personalization			
	Users rarely log in or make use of personalization options.	Provide and consider incentives of a user profile for (country and language specific) personalization and collaboration.	6.2.4

Table 8.1 Outcomes and recommendation from country and language specific logging

The results lead to the conclusion that sessions from different countries show significant variances with regard to the investigated variables. In particular, interface and result representation related interactions have the strongest impact on differences. Therefore, digital libraries need to consider usage patterns and requirements from users with different country and language backgrounds.

8.2 COMPLEMENTARY STUDIES AND FUTURE WORK

Based on the results of this study, at least two main future research directions are identified. First, the results from this study can be further investigated, proven and validated through the in-depth analysis of individual observations. This can be achieved by additional correlations between the variables or by complementary qualitative methods. To explain the non-usage of certain features such as the user profile, questionnaires or interviews would provide an insight into user motivations. The correlation of variables such as the interface language change with further session parameters reveals the impact of native language interfaces on usage patterns.

Second, the proposed country and language specific logging approach can be used for the evaluation and comparison of other multilingual digital libraries. Some of the observed interaction patterns might result from the system design and restrictions. Through the

comparison of search or browsing dominated systems, it can be examined to what extent users (from different countries) are influenced by functionalities.

As an example, three possible future study directions including further research questions are briefly described in the following paragraphs.

8.2.1 THE IMPACT OF THE INTERFACE LANGUAGE (CHANGE)

While the majority of digital libraries already offer a multilingual interface, previous research in line with this study has shown that most users do not switch the interface language but rather make use of the default (English) setting. On the other hand, everyday portals or applications (e.g. the search engine or browser) are customized to the native language. Independent of the application and implementation, the (mis-)understanding of the interface language (change) remains a challenge for system designers.

This study analyzed the occurrence of interface language changes as well as preferences for certain interface language change types and for native languages. So far, these variables were considered isolated from the session information. Since interface issues have been identified as high impact indicators for country and language specific interactions, it should be further investigated whether the change of the interface language has an influence on usage patterns. To evaluate the impact of native language accesses, sessions with and without interface language changes need to be compared regarding the duration of sessions, number of queries per sessions, the use of language facets as well as result interactions.

As an example, table 8.2 correlates the two variables Interface Language Change (LC) and Usage of Language Facet (LF) for German sessions. The first row contains the number of sessions with an interface language change (22,897 in total) either with language result refinement (14,616) or without (8,281). The second row displays the number of sessions without an interface language change (111,413 in total) either with language refinement (31,354) or without (80,059). The correlation shows that sessions without an interface language change and result refinement are prevalent. For sessions containing an interface language change, a higher usage of the language facet is observed. Consequently a relationship between the interface language selection and a higher usage of result language refinement is assumed. Similarly, other variables can be correlated to investigate country and language specific activity.

	Usage of Language Facet	No Usage of Language Facet
Interface Language Change	14,616	8,281
No Interface Language Change	31,354	80,059

Table 8.2 Correlation between interface language change and usage of facets for German sessions

The investigation of the correlation between the interface and query language will be of further interest. Previous studies of The European Library (TEL) query logs have shown that users sometimes tend to switch the interface language according to the query language (Gaede et al., 2011). Either the interface language change correlates with the query language change or users repeated the same query using different interface languages. Table 8.3 gives an example of a search session where a user types in translations of the original query and switches the interface language accordingly. The initial German query "fortfaitierung" (financial transaction) was submitted under the German interface. Instead of reformulating the query, the user changes the interface language to English and submits the English translation "fortaiting". Finally he changes the interface language back to German and types in the initial query again.

Interface Language	Query
DE	fortfaitierung
EN	fortfaiting
DE	fortfaitierung

Table 8.3 Correlation between interface language and query language

Another example shows a session where the user switches the interface language while submitting the same query again and again (table 8.4). In both examples, the user switches back to the initial interface language after trying other options. Based on these findings it could be assumed that users believe in a relationship between the interface language change and the language of their search.

Interface Language	Query
EN	bartolomeo bosco
IT	bartolomeo bosco
DE	bartolomeo bosco
EN	bartolomeo bosco

Table 8.4 Interface language change with language independent query

From the analysis of interface language changes of three systems – browser, Google and Europeana – contradictory results were derived. Additional research needs to be conducted to answer the questions why users change or do not change the interface language, whether users

understand the function of an interface language change as well as whether they prefer automatic interface language changes based on geo-location information or user-triggered changes.

The investigation of the emerging research questions requires the input of qualitative studies challenging the observed usage patterns.

8.2.2 THE IMPACT OF NATIVE CONTENT / SYSTEM REQUIREMENTS

The results of this study suggest that users are primarily interested in native language content, but not in native country content. Furthermore, the analysis has shown that it needs to be considered to what extent users have the possibility to access content in their native language. For Europeana, so-called content-rich and content-poor countries were identified, showing that users from these groups act differently when selecting native content. For the selection of native language facets as well as native language objects, the groups differ. Based on this observation, it can be assumed that users prefer native language to foreign objects. Accordingly, differences in behavior are caused by the available native content and system constraints. Consequently, future research needs to address the question to what extent the available native content influences users' interactions and respectively preferences. In order to study differences between countries with regard to result interactions and preferences, equivalent preconditions should be provided. Only if the investigated countries and language groups have access to the same amount of native content, a reliable statement concerning the preference of native content can be made.

Furthermore, the usage and role of foreign languages for the inspection of relevant results is an important issue when dealing with multilingual content. It needs to be determined to what extent users make use of foreign languages and in particular of English as secondary language during the search process. Consequently, it can be questioned if English metadata translations are a useful addition to the original object descriptions. Similarly, more research is required with regard to the usage and usefulness of translation services.

8.2.3 MULTILINGUAL QUERY ANALYSIS

In this study, a small sample set of 200 queries from Germany and France were manually assigned to categories and languages. The qualitative analysis suggests variations in the query content for each country. A larger data set including queries from all observed countries is required to investigate if users from different language and cultural backgrounds have different information needs. Another interesting comparison would involve country or culture specific differences between queries from countries with the same official language (e.g. Germany and Austria or GB and US). A third comparison could be made between frequently and non-frequently spoken languages. It could be assumed that smaller language groups use English more often as secondary language to find relevant results.

8.3 CONTRIBUTIONS

Based on findings from previous studies dealing with multilingual information access in digital libraries, this dissertation aims at providing a detailed picture of user behavior across countries, determining if country or language groups show similar patterns in information system interactions and if so, what characterizes them. The study considered language issues in digital libraries from the user, system and content perspective with respect to three main research questions:

RQ1: Which variables in log files can be leveraged to study the user's country and language context?

RQ2: Does usage data indicate country or language specific interaction patterns?

> **H_0**: Sessions from different countries and language backgrounds show the same interactions.
>
> **H_1**: Country and language level differences exist between sessions.

RQ3: Which variables gathered by log files uncover significant country and language specific differences in user interactions?

In chapter 5, a country and language specific logging approach was introduced. Based on a set of explicit and implicit indicators, the Europeana Language Logger (ELL) and an appropriate

analysis tool were developed and served as instrument for the study of country and language level differences within Europeana. The ELL combines transaction log data with additional clickstream data capturing application states providing a more detailed picture of user-system interactions. Europeana usage data from ten months containing 100,443,908 page views was collected and processed for further analysis. Sessions were clustered with regard to their origin and statistics gathered to identify country or language specific interaction patterns.

Chapter 6 presents the analysis of 1,071,872 sessions (after data cleaning and bot usage removal) from 21 countries. In total, 20 variables were investigated with regard to variances between country groups. For all variables and in particular for interface and result related interactions, significant differences exist between the countries. Most differences were observed for the usage and preference of native language interfaces as well as for the refinement and selection of native language content. Based on the available content within Europeana, a differentiation of content-rich and content-poor countries was proposed. Language or country of origin do not have a similar impact on search and browsing patterns, however the content of queries seems to be country specific – especially in a cultural heritage portal like Europeana. The duration of sessions and the number of unique queries per session also varies slightly between the country groups – possibly depending on the nature of the content (i.e. language) available for the user.

The influence of single variables on country attributes is discussed and high and low impact variables for country and language specific logging are suggested. From the 20 variables, 11 are classified as high impact indicators for country or language level differences in interactions. The strongest distinctive features are the usage of the Europeana interface language (change) as well as the usage of (native) language facets and content.

Based on the identified country characteristics, profiles were designed and graphically presented in chapter 7. A comparison is drawn between an individual country (here: France) to a median country profile, content-rich and content-poor countries (medians to individual profiles) and English and non-English countries (medians to individual profiles).

The study concludes with a summary of outcomes and recommendations for the implementation of multilingual access to digital libraries (chapter 8).

The dissertation contributes to the identification and understanding of country and language specific usage patterns. It has been demonstrated that sessions from different countries show significant variances with regard to the investigated variables. The methodology and analysis developed in this thesis generates insights that can lead to future research dealing with single

aspects in more detail. Based on results derived from the log analysis, further research should focus on the examination, explanation and validation of the observed patterns especially with regard to the impact of the interface language (change) and native language content available.

The limitations of this dissertation are mainly related to the focus on a single system. As every case study, the reported results are limited to the source of data (in this case the Europeana system) as it was present during the data gathering period. However, the proposed logging format can easily be adapted to other systems. The comparison of two or more systems helps identifying system overlapping behavior as well as system specific interactions.

Similarly, the usage of a single analysis method like log file analysis comes with advantages as well as challenges. Log file data is a rich source to show how users are interacting with a particular system. While the analysis of usage data provides insights into what is happening, it does not explain why this is the case. The interpretation of log files always lacks the user's background information, intentions and goals. For this study, no information about the actual users' language skills and preferences were gathered for the comparison with the log file data. Nevertheless, for the identification of country and language specific interaction patterns, individual users' information is peripheral.

This dissertation has contributed to the field of information behavior studies within digital library research. The special focus on language or country specific interactions provides a basis for the investigation and evaluation of multilingual digital library usage. The work needs to be continued with respect to open issues as well as the implementation of cross-lingual search functionalities that have not been part of this study. Only if truly multilingual systems are provided, a complete analysis of multilingual access can be conducted.

REFERENCES

Aggarwal, N., Buitelaar P. (2012). Query Expansion Using Wikipedia and Dbpedia. In Forner, P., Karlgren, J., Womser-Hacker, C (Eds.), *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*. Rome, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-CHiC-AggarwalEt2012.pdf

Agosti, M., Angelaki, G., Coppotelli, T., Di Nunzio, G. M. (2007). Analysing HTTP Logs of a European DL Initiative to Maximize Usage and Usability. In Goh, D. H.-L., Cao, T.H., Solvberg, I.T., Rasmussen, E. (Eds.), *Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers* (Lecture Notes in Computer Science, Vol. 4822, pp. 35-44). Berlin, Heidelberg: Springer.

Agosti, M., Coppotelli, T., Di Nunzio, G.M., Ferro, N., And Van Der Meulen, E. (2007a). A Study of Web Logs for Personalizing the Multilingual Information Access to The European Library. In Christodoulakis, S. (Ed.), *Workshop on Cross-Media and Personalized Learning Applications on top of Digital Libraries (LADL 2007), 11th European Conference on Research and Avanced Technology of Digital Libraries (ECDL 2007)* (Lecture Notes in Computer Science, Vol. 4675, pp. 19-28). Berlin, Heidelberg: Springer.

Agosti, M., Crivellari, F., Deambrosis, G. Ferro, N., Gäde M., Petras V., Stiller J. (2009). Report on User Preferences and Information Retrieval Scenarios for Multilingual Access in Europeana - D2.1.1 (EuropeanaConnect). Retrieved from http://www.europeanaconnect.eu/documents/D2.1.1_eConnect_Report_User_Preferences_MLIA_v1.0_20091222.zip

Agosti, M., Crivellari, F., Di Nunzio, G.M., Ioannidis, Y., Stamatogiannakis, E., Triantafyllidi, M.L., Vayanou, M. (2009). Report on Search Engines and HTTP Log Analysis - D5.2 (TELplus project). Retrieved from http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/documents/TELplus_D5.2_15102009.pdf

Agosti, M., Crivellari, F., Di Nunzio, G.M., Ioannidis, Y., Stamatogiannakis, E., Triantafyllidi, M.L., Vayanou, M. (2009a). Searching and Browsing Digital Library Catalogues: A Combined Log Analysis for the European Library. In Agosti, M., Esposito, F., Thanos, C. (Eds.), *Fifth Italian Research Conference on Digital Libraries (IRCDL 2009). A Conference of the DELOS Association and the Department of Information Engineering of the University of Padova* (pp. 120-135). Padova, Italy: DELOS Association.

Agosti, M., Crivellari, F., Di Nunzio G.M., Gabrielli S. (2010). Understanding User Requirements and Preferences for a Digital Library Web Portal. *International Journal on Digital Libraries, 11*(4), 225-238.

Agosti, M., Crivellari, F., Di Nunzio, G.M. (2012). Web Log Analysis: A Review of a Decade of Studies about Information Acquisition, Inspection and Interpretation of User Interaction. *Data Mining*

and Knowledge Discovery, 24(3), 663-696.

Amato, G., Cigarrán, J., Gonzalo, J., Peters, C., Savino, P. (2007). Multimatch-Multilingual/Multimedia Access to Cultural Heritage. In Kovacs, L., Fuhr, N., Meghini, C. (Eds.), *Research and Advanced Technology for Digital Libraries, 11th European Conference, ECDL 2007* (Lecture Notes of Computer Science, Vol. 4675, pp. 505-508). Berlin, Heidelberg: Springer.

Angelaki, G. (2007). Interim Report on Usability Developments in The European Library - M1.4 (EDLproject). Retrieved from http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/M1.4_Interim%20Report%20on%20Usage%20and%20Usability.pdf

Aula, A., Kellar, M. (2009). Multilingual Search Strategies. In Olsen Jr., D. R., Arthur, R.B., Hinckley, K., Ringel Morris, M. Hudson, S.E., Greenberg, S. (Eds.), *CHI '09 Extended Abstracts on Human Factors in Computing Systems* (pp. 3865-3870). Boston, USA: ACM.

Baeza-Yates, R. (2005). Applications of Web Query Mining. In Losada, D. E., Fernández-Luna, J.M. (Eds.), *Advances in Information Retrieval, 27th European Conference on IR Research (ECIR 2005)* (Lecture Notes in Computer Science, Vol. 7814, pp. 7-22). Berlin, Heidelberg: Springer.

Baeza-Yates, R., Calderon-Benavides, Gonzalez-Caro, C. (2006). The Intention Behind Web Queries. In Crestani, F., Ferragina, P., Sanderson, M (Eds.), *String Processing and Information Retrieval. Proceedings of the 13th International Conference (SPIRE 2006)* (Lecture Notes in Computer Science, Vol. 4209, pp. 98-109). Berlin, Heidelberg: Springer.

Beitzel, S. M., Jensen, E.C., Lewis, D.D., Chowdhury, A., Frieder, O. (2007). Automatic Classification of Web queries using Very Large Unlabeled Query Logs. *ACM Transactions on Information Systems 25*(2), Article 9.

Beitzel, S. M., Jensen, E., Chowdhury, A. , Frieder, O.,Grossman, D. . (2007). Temporal Analysis of a Very Large Topically Categorized Web Query Log. *Journal of the American Society for Information Science and Technology, 58*(2), 166-178.

Berendt, B., Mobasher, B., Nakagawa, M., Spiliopoulou, M. (2003). The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis. In Zaïane, O. R., Srivastava, J., Spiliopoulou, M., Masand, B. (Eds.), *WEBKDD 2002 - Mining Web Data for Discovering Usage Patterns and Profiles. Proceedings of the 4th International Workshop* (Lecture Notes in Computer Science, Vol. 2703, pp. 159-179). Edmonton, Canada: Springer.

Berendt, B., Kralisch, A. (2009). A User-Centric Approach to Identifying Best Deployment Strategies for Language Tools: The Impact of Content and Access Language on Web User Behaviour and Attitudes. *Information Retrieval, 12*(3), 380-399.

Berenson, M. L., Levine, D. M., .Krehbiel, T. C., Stephan, D. F. (2012). *Basic Business Statistics: Concepts and Applications* (12 ed.). New Jersey: Pearson.

Bernardi, R., Calvanese, D., Dini, L., Di Tomaso, V., Frasnelli, E., Kugler, U., Plank B. (2006). Multilingual Search in Libraries. The Case-Study of the Free University of Bozen-Bolzano. In Calzolari, N., Choukri, K., Gangemi, A., Maegaard, B., Mariani, J., Odijk, Merelbeke, J., Tapias, D. (Eds.), *5th International Conference on Language Resources and Evaluation (LREC 2006)* (pp. 2287-2290). Torino, Italy: Free University of Bozen-Bolzano.

Bernardi, R., Buoso P., Schiller, A., Gobbetti, D. (2008). Simple Interface - With Report Of Usability And Accessibility (CACAO project). Retrieved from http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D4.1.pdf

Bernardi, R., Balestrier, M., Bosca, A., Dini, L., Gobbetti, D., Segond, F. (2009). CACAO System: An Overview. In Bernardi, R., Gottfried, B., Segond, F., Zaihrayeu. I. (Eds.), *Advanced Technologies for Digital Libraries. International Workshop on NLP4DL (AT4DL 2009)* (Lecture Notes in Computer Science, Vol. 6699, pp. 61-65). Berlin, Heidelberg:Springer.

Berners-Lee, T., Fischetti, M. (1999). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* Britain: Orion Business.

Bilal, D., Bachir I. (2007). Children's Interaction with Cross-Cultural and Multilingual Digital Libraries I: Understanding Interface Design Representations. *Information Processing & Management, 43*(1), 47-64.

Bilal, D., Bachir I. (2007a). Children's Interaction with Cross-Cultural and Multilingual Digital Libraries II: Information Seeking, Success, and Affective Experience. *Information Processing & Management, 43*(1), 65-80.

Bilal, D., & Kirby, J. (2002). Differences and Similarities in Information Seeking: Children and Adults as Web Users. *Information Processing & Management, 38*(5), 649-670.

Borgman, C. L. (1989). All Users of Information Retrieval Systems are not Created Equal: An Exploration into Individual Differences. *Information Processing & Management, 25*(3), 237-251.

Borgman, C. L. (1997). Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries, Or How Do We Exchange Data in 400 Languages? *D-Lib Magazine 3*(6). Retrieved from http://dlib.org/dlib/june97/06borgman.html

Borgman, C. L., Rasmussen, E. (2005). Usability of Digital Libraries in a Multicultural Environment. In Theng, Y.-L., Foo, S. (Eds.), *Design and Usability of Digital Libraries: Case Studies in the Asia Pacific* (pp. 270 - 284). Hershey, PA: IGI Global.

Bosca, A., Dini, L. (2010). Language Identification Strategies for Cross Language Information Retrieval. In Braschler, M., Harman, D., Pianta, E. (Eds.), *CLEF 2010 LABs and Workshops, Notebook Papers*. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86374/CLEF2010wn-LogCLEF-BoscaEt2010.pdf.

Bourges-Waldegg, P., Scrivener, S.A.R. (1998). Meaning, the Central Issue in Cross-Cultural HCI Design. *Interacting with Computers, 9*(3), 287-309.

Braschler, M., Ferro, N. (2007). Adding Multilingual Information Access to The European Library. In Thanos, C., Borri, F., Candela, L. (Eds.), *Digital Libraries: Research and Development* (Lecture Notes in Computer Science, Vol. 4877, pp. 218-227). Berlin, Heidelberg: Springer.

Broder, A. (2002). A Taxonomy of Web Search. *SIGIR Forum, 36*(2), 3-10.

Bryan-Kinns, N., Blandford, A. (2000). A Survey of User Studies for Digital Libraries. *RIDL Working Paper.* Retrieved from http://www.uclic.ucl.ac.uk/annb/docs/DLuser.pdf

Buchanan, T., Paine, C., Joinson, A.J., Reips, U.-D. (2007). Development of Measures of Online Privacy Concern and Protection for Use on the Internet. *Journal of the American Society for Information*

Science and Technology, 58(2), 157-165.

Bucklin, R. E., Lattin, J. M., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J. D. C., Mela, C., Montgomery, A., Steckel, J. (2002). Choice and the Internet: From Clickstream to Research Stream. *Marketing Letters, 13*(3), 245-258.

Budzise-Weaver, T., Chen, J., Mitchell, M. (2012). Collaboration and Crowdsourcing: The Cases of Multilingual Digital Libraries. *Electronic Library, 30*(2), 220 - 232.

Burby, J. Brown, A. (2007). Web Analytics Definitions (Web Analytics Association). Retrieved from http://www.digitalanalyticsassociation.org/Files/PDF_standards/WebAnalyticsDefinitionsVol1.pdf

Burton, M., Walther, J. (2001). A Survey of Web Log Data and their Application in Use-based DesignSystem Sciences 2001. PROCEEDINGS OF THE ANNUAL HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES. Washington, USA: EEE Computer Society. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.135.8755&rep=rep1&type=pdf.

Caffo, R., Hagedorn-Saupe, M., Zakrajšek, F. (2008). Handbook on Cultural Web User Interaction (Minverca EC). Retrieved from http://www.minervaeurope.org/publications/handbookwebusers-firstdraft-june08.pdf

Caidi, N., Komlodi, A. (2003). Digital Libraries Across Cultures: Design and Usability Issues Outcomes of the "Cross-Cultural Usability for Digital Libraries" Workshop at JCDL '03. *SIGIR Forum, 37*(2), 62-64.

Ceccarelli, D., Gordea, S., Lucchese, C., Nardini, F.M., Tolomei, G. (2011). Improving Europeana Search Experience Using Query Logs. In Gradmann, S., Borri, F., Meghini, C., Schuldt, H. (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries* (Lecture Notes in Computer Science, Vol. 6966, pp. 384-395). Berlin, Heidelberg: Springer.

Chen, J., Bao, Y. (2009). Information Access across Languages on the Web: From Search Engines to Digital Libraries. *Proceedings of the American Society for Information Science and Technology, 46*(1), 1-14.

Chen, M. S., Park, J. S., Yu, P. S. (1998). Efficient Data Mining for Path Traversal Patterns. *IEEE Transactions on Knowledge and Data Engineering, 10*(2), 209 - 221.

Chowdhury, S., Landoni, M., Gibb, F. (2006). Usability and Impact of Digital Libraries: A Review. *Online Information Review, 30*(6), 656 – 680.

Chu, P., Jozsa, E., Komlodi, A., Hercegfi, K. (2012). An Exploratory Study on Search Behavior in Different Languages. In Kamps, J., Kraaij, W., Fuhr, N. (Eds.), *Proceedings of the 4th Information Interaction in Context Symposium* (pp. 318-321). New York, NY: ACM.

CIBER Research Ltd (2013). Europeana 2012-2013: Usage and Performance update. Growth, Change and Impact; Mobile devices, Stickness, Loyalty, Social Media, Virtual Exhibitions and Methodology. Retrieved from http://ciber-research.eu/download/20130623-Europeana_2013_usage_and_performance_update.pdf

Clark, D. J., Nicholas, D., Rowlands, I. (2011). Publishable Report on Best Practice and How Users are

Using the Europeana Service - D3.1.3 (EuropeanaConnect project). Retrieved from http://www.europeanaconnect.eu/documents/D3.1.3_eConnect_LogAnalysisReport_v1.0.pdf

Clavel-Merrin, G., Žumer, M., Landry, P. (2006). Scenarios for Multilingual Access - D3.6 (TEL-ME-MOR project). Retrieved from http://www.nuk.uni-lj.si/telmemor/docs/D3.6_Multilingual_scenarios.pdf

Clavel-Merrin, G., Pisanski, J., Žumer, M.,. (2008). Multilingual Achievements in EDL and Remaining Challenges - D2.1 (TEL project). Retrieved from http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/D2%201final.pdf

Cleverdon, C. W. (1970). The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages. *Technical Report Part 3*. Retrieved from http://dspace.lib.cranfield.ac.uk/handle/1826/967

Clifton, B. (2010). *Advanced Web Metrics with Google Analytics, 2nd Edition*. Indianapolis, Indiana: SYBEX Inc.

Clough, P., Sanderson, M. (2006). User Experiments with the Eurovision Cross-Language Image Retrieval System. *Journal of the American Society for Information Science and Technology, 57*(5), 697-708.

Clough, P., Eleta, I. (2010). Investigating Language Skills and Field of Knowledge on Multilingual Information Access in Digital Libraries. *International Journal of Digital Library Systems (IJDLS), 1*(1), 89-103.

Cooley, R., Tan, P.-N., Srivastava, J. (2000). Discovery of Interesting Usage Patterns from Web Data. In Masand, B. M., Spiliopoulou, M. (Eds.), *Web Usage Analysis and User Profiling. International WEBKDD'99 Workshop* (Lecture Notes in Computer Science, Vol. 1836, pp. 163-182). Berlin, Heidelberg: Springer.

Cooper, M. D. (2001). Usage Patterns of a Web-Based Library Catalog. *Journal of the American Society for Information Science and Technology, 52*(2), 137-148.

Cousins, J. (2006). The European Library – Pushing the Boundaries of Usability. *The Electronic Library, 24*(4), 434-444.

Covey, D. T. (2002). Usage and Usability Assessment: Library Practices and Concerns. *Digital Library Federation, Council on Library and Information Resources*. Retrieved from http://www.clir.org/pubs/reports/pub105/pub105.pdf

Dekkers, M., Gradmann, S., Meghini, C. (2009). Europeana Outline Functional Specification for Development of an Operational European Digital Library (Europeana v1.0 project). Retrieved from http://abm.ylm.se/europeanalocal/pdf/EuropeanaOutline08.pdf

Deutscher, G. (2010). *Through the Language Glass: Why the World Looks Different in Other Languages*. London, UK: Henry Holt and Company.

Di Nunzio, G. M. (2008). "Interactive" Undergraduate Students: UNIPD at iCLEF 2008. In Peters, C. (Ed.), *Results of the CLEF 2008 Cross-Language System Evaluation Campaign*. Aarhus, Denmark. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86371/CLEF2008wn-iCLEF-DiNunzio2008.pdf.

Directorate-General Information Society and Media. (2011). User Language Preferences Online (Analytics Report) Vol. 313. *Flash Eurobarometer*. Retrieved from http://ec.europa.eu/public_opinion/flash/fl_313_en.pdf

Diekema, A. R. (2012). Multilinguality in the Digital Library: A Review. *The Electronic Library, 30*(2), 165-181.

Dobreva, M., McCulloch, E., Birrell, D., Unal, Y., Feliciati, P. (2010). Digital Natives and Specialised Digital Libraries: A Study of Europeana users. In Kurbanoğlu, S., Al, U., Erdoğan, P. L., Yaşar Tonta, Y., Uçak, N. (Eds.), *Technological Convergence and Social Networks in Information Management. Second International Symposium on Information Management in a Changing World (IMCW 2010)* (Communications in Computer and Information Science, Vol. 96, pp. 45-60). Berlin, Heidelberg: Springer.

Dobreva, M., Chowdhury, S. (2010). A User-Centric Evaluation of The Europeana Digital Library. In Chowdhury, G., Khoo, C., Hunter, J. (Eds.), *The Role of Digital Libraries in a Time of Global Change, 12th International Conference on Asia-Pacific Digital Libraries* (Lecture Notes in Computer Science, Vol. 6102, pp. 148-157). Berlin, Heidelberg: Springer.

Dobreva, M., McCulloch, E., Birrell, D., Feliciati, P., Ruthven, I., Sykes, J., Unal, Y. (2010a). User and Functional Testing-Final report (Europeana v1.0 project). Retrieved from http://pro.europeana.eu/c/document_library/get_file?uuid=1c25ae28-9457-4b0f-be62-654a7cf6c5b7&groupId=10602

Dobreva, M., O'Dwyer, A., Feliciati, P. (2012). *User Studies for Digital Library Development*. London, UK: Facet Publishing.

Doran, D., Gokhale, S.S. (2011). Web Robot Detection Techniques: Overview and Limitations. *Data Mining and Knowledge Discovery, 22*(1-2), 183-210.

Druin, A., Bederson, B. B., Hourcade, J.P., Sherman, L., Revelle, G., Platner, M., Weng, S. (2001). Designing a Digital Library for Young Children: An Intergenerational Partnership. In Fox, E. A., Borgman, C.L. (Eds.), *Proceedings of Joint Conference on Digital Libraries (JCDL 2001)* (pp. 398 - 405). New York, NY: ACM.

Europeana Foundation. (2011). Europeana Strategic Plan 2011 - 2015 (Europeana Foundation). Retrieved from http://www.pro.europeana.eu/c/document_library/get_file?uuid=c4f19464-7504-44db-ac1e-3ddb78c922d7&groupId=10602

Fagni, T., Perego, R., Silvestri, F., Orlando, S. . (2006). Boosting the Performance of Web Search Engines: Caching and Prefetching Query Results by Exploiting Historical Usage Data. *ACM Transactions on Information Systems, 24*(1), 51-78.

Fang, W. (2007). Using Google Analytics for Improving Library Website Content and Design: A Case Study. *Library Philosophy and Practice (e-journal), Paper 121*. Retrieved from http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1121&context=libphilprac

Ferro, N., Masiero, I., Nicchio, M., Peruzzo, S., Silvello, G. (2012). Creation of the Experimental Collection for CHiC2012 – Task 6.2: (PROMISE project).

Ford, G., Gelderblom, H. (2003). The Effects of Culture on Performance Achieved Through the Use of Human Computer Interaction. In Eloff, J., Engelbrecht, A., Kotzé, P., Eloff, M. (Eds.),

Proceedings of the 2003 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on Enablement through Technology (pp. 218-230). Republic of South Africa: South African Institute for Computer Scientists and Information Technologists.

Ford, G., Kotzé, P. (2005). Designing Usable Interfaces with Cultural Dimensions. In Costabile, M. F., Paternò, F. (Eds.), *Human-Computer Interaction - INTERACT 2005, IFIP TC13 International Conference* (Lecture Notes in Computer Science, Vol. 3585, pp. 713-726). Berlin, Heidelberg: Springer.

Ford, G., Kotzé, P (2005a). Researching Culture and Usability – A Conceptual Model of Usability. In McEwan, T., Gulliksen, J., Benyon, D. (Eds.), *People and Computers XIX — The Bigger Picture, Proceedings of the British Human Computer Interaction Conference 2005* (pp. 317 - 333). London, UK: Springer.

Ford, N., Miller, D., Moss, N. (2001). The Role of Individual Differences in Internet Searching: An Empirical Study. *Journal of the American Society for Information Science and Technology, 52*(12), 1049-1066.

Franklin, B., Kryillidou, M., Plum, T., Tsakonas, G., Papatheodorou, C. (2009). From Usage to User: Library Metrics and Expectations for the Evaluation of Digital Libraries. In Tsakonas, G., Papatheodorou, C. (Eds.), *Evaluation of Digital Libraries: An Insight into Useful Aplications and Methods* (Chandos Information Professional Series, pp. 17-52). Great Abington, UK: Chandos.

Fresa, A. (2005). MICHAEL: Multilingual Inventory of Cultural Heritage in Europe (Michael project). Retrieved from http://www.michael-culture.eu/documents/fresaeva05.pdf

Frieseke, M., Gäde, M., Petras, V., Stiller, J. (2011). Interaction Patterns in Cultural Heritage Information Systems - Internal Report (PROMISE project).

Fuhr, N., Hansen, P., Mabe, M., Micsik, A., & Sølvberg, I. (2001). Digital Libraries: A Generic Classification and Evaluation Scheme. In Constantopoulos, P., Sølvberg, I. (Eds.), *Research and Advanced Technology for Digital Libraries. 5th European Conference (ECDL 2001)* (Lecture Notes in Computer Science, Vol. 2163, pp. 187-199). Berlin, Heidelberg: Springer.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Solvberg, I. (2007). Evaluation of Digital Libraries. *International Journal on Digital Libraries, 8*(1), 27-38.

Gäde, M., Petras, V., Stiller J. (2010). Which Log for Which Information? Gathering Multilingual Data from Different Log File Types. In Agosti, M. F., N.; Peters, C.; de Rijke, M.; Smeaton, A. (Eds.), *Multilingual and Multimodal Information Access Evaluation. Proceedings of the 2010 International Conference on Multilingual and Multimodal Information Access Evaluation: Cross-Language Evaluation Forum* (Lecture Notes in Computer Science, Vol. 6360, pp. 70-81). Berlin, Heidelberg: Springer.

Gäde, M., Stiller, J., Berendsen, R., Petras, V. (2011). Interface Language, User Language and Success Rates in the European Library. In Petras, V., Forner, P., Clough, P. D. (Eds.), *CLEF 2011 LABs and Workshops, Notebook Papers*. Amsterdam, Netherlands. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86377/CLEF2011wn-LogCLEF-GaedeEt2011.pdf.

Gäde, M., Stiller, J. (2011). Multilingual Interface Usage. In Griesbaum, J., Mandle, T., Womser-Hacker, C. (Eds.), *Information und Wissen: Global, Sozial und Frei? Proceedings des 12. Internationalen Symposiums für Informationswissenschaft.* (pp. 503-504). Boizenburg, Germany: Hülsbusch Verlag.

Gäde, M., Ferro, N., Paramita, M.L. (2011). CHiC 2011 – Cultural Heritage in CLEF: From Use Cases to Evaluation in Practice for Multilingual Information Access to Cultural Heritage. In Petras, V., Forner, P., Clough, P.D. (Eds.), CLEF 2011 LABs and Workshops, Notebook Papers. Amsterdam, Netherlands. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86377/CLEF2011wn-CHiC-GaedeEt2011.pdf.

Gandal, N., Shapiro, C. (2001). The Effect of Native Language on Internet Usage. Retrieved from http://arxiv.org/pdf/cs/0109009.pdf

Gey, F. C., Kando, N., Peters, C. (2005). Cross-Language Information Retrieval: The Way Ahead. *Information Processing & Management, 41*(3), 415-431.

Gey, F. C. Kando, N., Lin, C.-Y., Peters, C. (2006). New Directions in Multilingual Information Access. *SIGIR Forum, 40*(2), 31-39.

Ghorab, M. R., Leveling, J., Zhou, D., Jones, G.J.F., Wade, V. (2010). Identifying Common User Behaviour in Multilingual Search Logs. In Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments. Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments* (Lecture Notes in Computer Science, Vol. 6241, pp. 518-525). Berlin, Heidelberg: Springer.

Gonçalves, M. A., Fox, E. A., Watson, L. T., & Kipp, N. A. (2004). Streams, Structures, Spaces, Scenarios, Societies (5s): A Formal Model for Digital Libraries. *ACM Transactions on Information Systems, 22*(2), 270-312.

Gonzalo, J., Penas, A., Verdejo, F., Peters, C. (2008). Workshop on Best Practices for the Development of Multilingual Information Access Systems: The User Perspective – D3.2 (TrebleCLEF project). Retrieved from http://www.trebleclef.eu/getfile.php?id=249

Gravano, L., Hatzivassiloglou, V., Lichtenstein, R. (2003). Categorizing Web Queries According to Geographical Locality. In Grossman, D. A., Gravano, L., Zhai, C.X., Herzog, O., Evans, D.A. (Eds.), *Proceedings of the Twelfth International Conference on Information and Knowledge Management* (pp. 325-333). New York, NY: ACM.

Grefenstette, G., Nioche, J. (2000). Estimation of English and non-English Language Use on the WWW. In Mariani, J.-J., Harman, D. (Eds.), *Computer-Assisted Information Retrieval (Recherche d'Information et ses Applications). Proceedings of 6th International Conference* (pp. 237 - 246). New York, NY: ACM.

Greifeneder, E. (2012). *Does it Matter Where We Test?: Online User Studies in Digital Libraries in Natural Environments* (Dissertation), Humboldt-Universität zu Berlin

Grimes, C., Tang, D. , Russell, D. M. (2007). Query Logs alone are not Enough. *Query Log Analysis: Social and Technological Challenges Workshop held in line with the 16th World Wide Web Conference, WWW 2007*. Retrieved from

http://static.googleusercontent.com/media/research.google.com/de//pubs/archive/34431.pdf

Guldbæk Rasmussen, K., Iversen, R., Petersen, G. (2010). Personas Catalogue - M3.2.3 (EuropeanaConnect project). Retrieved from http://www.europeanaconnect.eu/documents/M3.2.3_eConnect_PersonasCatalogue_v1.0_20091228.pdf

Halder, S., Ray, A., & Chakrabarty, P. K. (2010). Gender Differences in Information Seeking Behavior in Three Universities in West Bengal, India. *The International Information & Library Review, 42*(4), 242-251.

Hall, E. T., Hall, M.R. (1990). *Understanding Cultural Differences - Germans, French and Americans*. Bosten, USA: Intercultural Press, Incorporated.

Haselhuber, J. (2012). Mehrsprachigkeit in der Europäischen Union: Eine Analyse der EU-Sprachpolitik mit besonderem Fokus auf Deutschland *Duisburger Arbeiten zur Sprach- und Literaturwissenschaften* (Vol. 92). Frankfurt am Main: Lang.

Hawkey, K. (2008). Privacy Concerns for Web Logging Data. In Jansen, B. J., Spink, A., Taksai, I. (Eds.), *Handbook of Research on Web Log Analysis.* Hershey, USA: Information Science Reference.

He, D., Göker, A. (2000). Detecting Session Boundaries from Web User Logs. Retrieved from http://static.googleusercontent.com/media/research.google.com/de//pubs/archive/34431.pdf

He, D., Göker, A., Harper, D.J. (2002). Combining Evidence for Automatic Web Session Identification. *Information Processing & Management, 38*(5), 727-742.

Hearst, M. (2009). *Search User Interfaces*. New York, NY: Cambridge University Press.

Hofmann, K., Rijke, M. de, B. Huurnink, B., Meij, E. J. (2009). A Semantic Perspective on Query Log Analysis. In Peters, C. (Ed.), *Results of the CLEF 2009 Cross-Language System Evaluation Campaign*. Corfu, Greece. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/85150/CLEF2009wn-LogCLEF-HofmannEt2009.pdf.

Hofstede, G. (1983). National Cultures in Four Dimensions: A Research-Based Theory of Cultural Differences among Nations. *International Studies of Management & Organization, 13*(1-2), 46-74.

Hofstede, G., Hofstede, G.J., Minkov, M. (2010). *Cultures and Organizations: Software of the Mind, Intercultural Cooperation and Its Importance fir Survival (Third Edition)*. USA: McGraw-Hill Professional.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6*(2), 65 - 70.

Hsieh, H. C., Holland, R., Young, M. (2009). A Theoretical Model for Cross-Cultural Web Design. In Kurosu, M. (Ed.), *Human Centered Design. Proceedings of the 1st International Conference on Human Centered Design: Held as Part of HCI International 2009* (Lecture Notes in Computer Science, Vol. 5619, pp. 712-721). Berlin, Heidelberg Springer.

Huang, X., Peng, F., An, A., Schuurmans, D. (2004). Dynamic Web Log Session Identification with Statistical Language Models. *Journal of the American Society for Information Science and Technology, 55*(14), 1290 - 1303.

Huntington, P., Nicholas, D., Jamali, H. R. (2008). Website Usage Metrics: A Re-Assessment of Session

Data. *Information Processing & Management, 44*(1), 358-372.

Hutchinson, H. B., Rose, A., Bederson, B., Weeks, A. C., Druin, A. . (2005). The International Children's Digital Library: A Case Study in Designing for a Multi-Lingual, Multicultural, Multi-Generational Audience. *Information Technology and Libraries, 24*(1), 4-12.

Huynh, T., Miller, J. (2009). Empirical Observations on the Session Timeout Threshold. *Information Processing & Management, 45*(5), 513-528.

Ingwersen, P., Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. New York, NY: Springer.

IRN Research. (2009). EUROPEANA - Online Visitor Survey Research Report Version 3. Retrieved from http://pro.europeana.eu/c/document_library/get_file?uuid=e165f7f8-981a-436b-8179-d27ec952b8aa&groupId=10602

IRN Research. (2011). EUROPEANA - Online Visitor Survey Research Report Version 3. Retrieved from http://pro.europeana.eu/c/document_library/get_file?uuid=334beac7-7fc2-4a4e-ba23-4dcc1450382d&groupId=10602

Ishida, R., Miller, S. K. . (2006). Localization vs. Internationalization. *W3C i18n Activity*. Retrieved from http://www.w3.org/International/questions/qa-i18n.en.php

Jansen, B. J., Spink, A., Saracevic, T. (2000). Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web. *Information Processing & Management, 36*(2), 207-227.

Jansen, B. J., Spink, A. (2005). An Analysis of Web Searching by European AlltheWeb.com Users. *Information Processing & Management, 41*(2), 361-381.

Jansen, B. J. (2006). Search Log Analysis: What It Is, What's Been Done, How To Do It. *Library & Information Science Research, 28*(3), 407–432.

Jansen, B. J., Spink, A., Blakely, C., Koshman, S. (2007). Defining a Session on Web Search Engines: Research Articles. *Journal of the American Society for Information Science and Technology, 58*(6), 862-871.

Jansen, B. J. (2009). The Methodology of Search Log Analysis. In Jansen, B. J., Spink, A., Taksai, I. (Eds.), *Handbook of Research on Web Log Analysis* (pp. 100-123). Hershey, USA: Information Science Reference.

Jansen, B. J., Booth, D. L., Spink, A. (2007). Determining the User Intent of Web Search Engine Queries. In Fridman Noy, N., Alani, H., Stumme, G., Mika, P., Sure, Y., Vrandecic, D. (Eds.), *Proceedings of the 16th International Conference on World Wide Web* (pp. 1149-1150). New York, NY: ACM.

Jansen, B. J., Pooch, U. (2001). A Review of Web Searching Studies and a Framework for Future Research. *Journal of the American Society for Information Science and Technology, 52*(3), 235-246.

Jensen, C., Potts, C., Jensen, C. (2005). Privacy Practices of Internet Users: Self-Reports versus Observed Behavior. *International Journal of Human-Computer Studies, 63*(1-2), 203-227.

Jesper, S., Clough, P., Hall, M. (2013). Regional Effects on Query Reformulation Patterns. In Aalberg, T., Papatheodorou, C., Dobreva, M., Tsakonas, G., Farrugia, C.J (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)* (Lecture Notes in Computer Science, Vol. 8092, pp. 382 -

385). Berlin, Heidelberg Springer.

Joachims, T. (2002). Optimizing Search Engines using Clickthrough Data. In Zaïane, O. R., Goebel, R., Hand, D., Keim, D., Ng, R. (Eds.), *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 133-142). New York, NY: ACM.

Jones, G. J. F., Zhang, Y., Fantino, F., Newman, E., Debole, F. (2007). Multilingual Search for Cultural Heritage Archives by Combining Multiple Translation Resources. In van den Bosch, A., Grover, C., Sporleder, C (Eds.), *ACL Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)* (pp. 81-88). Prague, Czech Republic: Association for Computational Linguistics.

Jones, M., Alony, I. (2007). The Cultural Impact of Information Systems – Through the Eyes of Hofstede - A Critical Journey. In Cohen, E. B. (Eds.), *Information and Beyond Part 1. Issues in Informing Science and Information Technology (IISIT)* (pp. 407-419): California, USA :Informing Science Press.

Jones, S., Cunningham, S.J., McNab, R.J. (1998). An Analysis of Usage of a Digital Library. In Nikolaou, C., Stephanidis, C. (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries* (Lecture Notes in Computer Science, Vol. 1513, pp. 261-277). Berlin, Heidelberg: Springer.

Józsa, E., Köles, M., Komlódi, A., Hercegfi, K., Chu, P. (2012). Evaluation of Search Quality Differences and the Impact of Personality Styles in Native and Foreign Language Searching Tasks. In Fuhr, N., Kamps, J., Kraaij, W. (Eds.), *Proceedings of the 4th Information Interaction in Context Symposium* (pp. 310-313). New York, NY: ACM.

Kamps, J., Geva, S., Peters, C., Sakai, T., Trotman, A. & Voorhees, E. . (2009). Report on the SIGIR 2009 Workshop on the Future of Information Retrieval Evaluation. *SIGIR Forum, 43*(2).

Kani-Zabihi E., G. G., Chen, S.Y (2006). Digital Libraries: What Do Users Want? *Online Information Review, 30*(4), 395 -412.

Kaushik, A. (2007). *Web Analytics: An Hour A Day* (6th ed.). Indianapolis, Indiana: John Wiley & Sons.

Keegan, T. T., Cunningham, S.J. (2005). Language Preference in a Bi-Language Digital Library. In Marlino, M., Sumner, T., Shipman, F. (Eds.), *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries* (pp. 174-175). New York, NY: ACM.

Kellar, M., Watters, C. and Shepherd, M. A. (2006). Goal-Based Classification of Web Information Tasks. *Proceedings of the American Society for Information Science and Technology (ASIST)* (1, Vol. 43, pp. 1-22). Austin, TX: American Society for Information Science and Technology.

Keller, P., Oomen, J. (2010). Europeana Policy on User-generated Content - D4.1 (Europeana v1.0 project). Retrieved from pro.europeana.eu/documents/844813/851957/D1.4+UGC+policy.pdf

Kelly, D. (2009). Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends in Information Retrieval, 3*(1-2), 1-224.

Ketchen, D. J., Shook, C.L. (1998). The Application of Cluster Analysis. An Analysis and Critique. *Strategic Management Journal, 17*(6), 441-458.

Khoo, M., Pagano, J., Washington, A.L., Recker, M., Palmer, B., Donahue, R.A. (2008). Using Web Metrics to Analyze Digital Libraries. In Larsen, R., Paepcke, P., Borbinha, J., Naaman, M.

(Eds.), *Proceedings of the 8th ACM/IEEE-CS Joint Conference on Digital libraries (JCDL)* (pp. 375-384). New York, NY: ACM.

Klas, C. P., Fuhr, N., Kriewel, S., Albrechtsen, H., Tsakonas, G., Kapidakis, S., Papatheodorou, C., Hansen, P., Kovacs, L., Micsik, A., Jacob, E. (2006). An Experimental Framework for Comparative Digital Library Evaluation: The Logging Scheme. In Marchionini, G., Nelson, M.L., Marshall, C.C. (Eds.), *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)* (pp. 308-309). New York, NY: ACM.

Koch, T., Ardo, A., Golub, K. (2004). Browsing and Searching Behavior in the Renardus Web Service: A Study based on Log Analysis. In Dijken, C. v., Blasius, J., Durand, C. (Eds*.), Recent Developements and Applications in Social Research Methdology: Proceedings of the RC33 Sixth International Conference on Social Science Methodology*. Opladen: Budrich Verlag. Retrieved from http://opus.bath.ac.uk/14366/1/day-renardus-paper-v2.pdf.

Komlodi, A., Caidi, N., Wheeler, K. (2004). Cross-Cultural Usability of Digital Libraries. In Chen, Z., Chen, H., Miao, Q., Fu, Y., Fox, E.A., Lim, E.-P (Eds.), *Digital Libraries: International Collaboration and Cross-Fertilization, 7th International Conference on Asian Digital Libraries (ICADL 2004)* (pp. 584-593). Berlin, Heidelberg: Springer.

Kralisch, A. (2005). *The Impact of Culture and Language on the Use of the Internet: Empirical Analysis of Behaviour and Attitudes.* (Dissertation), Humboldt-Universität zu Berlin Germany.

Kralisch, A., Mandl, T. (2006). Barriers to Information Access across Languages on the Internet: Network and Language Effects. In Sprague, R. H., Laney, S., Robichaud, E. (Eds.), *Proceedings of the 39th Annual Hawaii International Conference on System Sciences* (pp. 54.52). Washington, USA: IEEE Computer Society.

Kürsten, J., Wilhelm, T., Richter, D., Eibl, M. (2012). Chemnitz at the CHiC Evaluation Lab 2012: Creating an Xtrieval Module for Semantic Enrichment. In Forner, P., Karlgren, J., Womser-Hacker, C (Eds.), *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*. Rome, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-CHiC-K%C3%BCrstenEt2012.pdf.

Kurth, M. (2003). The Limits and Limitations of Transaction Log Analysis. *Library Hi Tech, 11*(2), 98-104.

Lamb, R., Kling, R. (2003). Reconceptualizing Users as Social Actors in Information Systems Research. *MIS Quarterly, 27*(2), 197-235.

Lamm, K., Mandl, T., Koelle, R. (2010). Search Path Visualization and Session Performance Evaluation with Log Files. In Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments. Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments* (Lecture Notes in Computer Science, Vol. 6241, pp. 538-543). Berlin, Heidelberg: Springer.

Large, A., Moukdad, H. (2000). Multilingual Access to Web Resources: An Overview. *Electronic Library and Information Systems, 34*(1), 43-58.

Lau, T., Horvitz, E. (1999). Patterns of Search: Analyzing and Modeling Web Query Refinement. In Kay,

J. (Ed.), *Proceedings of the Seventh International Conference on User Modeling* (pp. 119-128). Secaucus, USA: Springer.

Lee, H. J. (2011). *Google Analytics for Digital Library Evaluation.* (Master), Oslo and Akershus University College of Applied Sciences Norway.

Lee, U., Liu, Z., Cho, J. (2005). Automatic Identification of User Goals in Web Search. In Ellis, A., Hagino, T (Eds.), *Proceedings of the 14th International Conference on World Wide Web* (pp. 391-400). New York, NY: ACM.

Lei, J. Z., Ghorbani, A. A. (2004). The Reconstruction of the Interleaved Sessions from a Server Log. In Tawfik, A. Y., Goodwin, A.D. (Eds.), *Advances in Artificial Intelligence. Procedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence* (Lecture Notes in Computer Science, Vol. 3060, pp. 133 - 145). Berlin, Heidelberg: Springer.

Leveling, J., Ghorab, R., Magdy, W., Jones, G. J. F., Wade, V. (2010). DCU-TCD@LogCLEF 2010: Re-Ranking Document Collections and Query Performance Estimation. In Braschler, M., Harman, D., Pianta, E. (Eds.), CLEF *2010 LABs and Workshops, Notebook Papers*. Padua, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86374/CLEF2010wn-LogCLEF-LevelingEt2010.pdf.

Levergood, B., Farrenkopf, S., Frasnelli, E. (2008). The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO project). In Greenberg, J. (Eds.), *Proceedings of the 2008 International Conference on Dublin Core and Metadata Applications* (pp. 191-196). Dublin, OH: Dublin Core Metadata Initiative.

Liew, C. L. (2009). Digital Library Research 1997-2007: Organisational and People Issues. *Journal of Documentation, 65*(2), 245-266.

Mandl, T., Agosti, M., Di Nunzio, G.M., Yeh, A., Mani, I., Doran, C., Schulz, J.M. (2010). LogCLEF 2009: The CLEF 2009 Multilingual Logfile Analysis Track Overview. In Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda G. (Eds.), *Multilingual Information Access Evaluation I. Text Retrieval Experiments. Proceedings of the 10th Cross-Language Evaluation Forum Conference on Multilingual Information Access Evaluation: Text Retrieval Experiments* (Lecture Notes in Computer Science, Vol. 6241, pp. 508-517). Berlin, Heidelberg: Springer.

Mandl, T., Di Nunzio, G.M., Schulz, J.M. (2010a). LogCLEF 2010: The CLEF 2010 Multilingual Logfile Analysis Track Overview. In Braschler, M., Harman, D., Pianta, E. (Eds.), CLEF 2010 LABs and Workshops, Notebook Papers (pp. 22-23). Padua, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86374/CLEF2010wn-LogCLEF-MandlEt2010.pdf.

Mane, L. (2009). Improving Full-Text Search in Printed Digital Libraries' Collections through Semantic and Multilingual Functionalities - Technologies Assessment & User Requirements - D3.2. (TELplus project).

Marascuilo, L. A. (1966). Large-Sample Multiple Comparisons. *Psychological Bulletin, 65*, 280-290.

Marchionini, G., Plaisant, C., Komlodi, A. (2003). The People in Digital Libraries: Multifaceted Approaches to Assessing Needs and Impact. In Bishop, A., Buttenfield, B., van House, N. (Eds.), *Digital Library Use: Social Practice in Design and Evaluation* (pp. 119-160). Cambridge: MIT

Press.

Marcus, A., Baumgartner, V.J. (2004). A Practical Set of Culture Dimensions for Global User-Interface Development. In Masoodian, M., Jones S., Rogers, B. (Eds.), *Computer Human Interaction. Proceedings of the 6th Asia Pacific Conference (APCHI 2004)* (Lecture Notes in Computer Science, Vol. 3101, pp. 252-261). Berlin, Heidelberg: Springer.

Marcus, A., Alexander, C. (2007). User Validation of Cultural Dimensions of a Website Design. In Aykin, N. (Ed.), *Usability and Internationalization. Global and Local User Interfaces. Proceedings of the 2nd International Conference on Usability and Internationalization* (Lecture Notes in Computer Science, Vol. 4560, pp. 160-167). Berlin, Heidelberg: Springer.

Marlow, J., Clough, P.D., Dance, K. (2007). Multilingual Needs of Cultural Heritage Web Site Visitors: A Case Study of Tate Online. In Trant, J., Bearman D. (Eds.), *Proceedings of the International Cultural Heritage Informatics Meeting (ICHIM07).* Toronto, Canada: Archives & Museum Informatics. Retrieved from http://www.archimuse.com/ichim07/papers/marlow/marlow.html.

Marlow, J., Clough, P.D., Recuero, J.C., Artiles, J. (2008a). Exploring the Effects of Language Skills on Multilingual Web Search. In Macdonald, C., Ounis, I., Plachouras, V., Ruthven, I., White, R.W. (Eds.), *Advances in Information Retrieval. Proceedings of 30th European Conference on IR Research (ECIR 2008)* (Lecture Notes in Computer Science, pp. 126-137). Berlin, Heidelberg: Springer.

Marlow, J., Clough, P.D., Ireson, N., Cigarran Recuero, J., Artiles, J., Debole, F. (2008b). The Multimatch Project: Multilingual/Multimedia Access To Cultural Heritage On The Web. In Trant, J., Bearman, D. (Eds.), *Museums and the Web*. Toronto: Archives & Museum Informatics. Retrieved from http://www.archimuse.com/mw2008/papers/marlow/marlow.html.

McGann, R. (2005). Study: Consumers Delete Cookies at Surprising Rate. *ClickZ Marketing News and Expert Advice*. Retrieved from http://www.clickz.com/clickz/news/1691871/study-consumers-delete-cookies-surprising-rate

Meij, E., Bron, M., Hollink, L., Huurnink, B., & Rijke, M. (2009). Learning Semantic Query Suggestions. In Bernstein, A., Karger, D. R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (Eds.), *The Semantic Web. Proceedings of the 8th International Semantic Web Conference (ISWC 2009)* (Lecture Notes in Computer Science, Vol. 5823, pp. 424-440). Berlin, Heidelberg: Springer.

Meiss, M., Duncan, J., Goncalves, B., Ramasco, J.J., Menczer, F. (2009). What's in a Session: Tracking Individual Behavior on the Web. In Cattuto, C., Ruffo, G., Menczer, F. (Eds.), *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia* (pp. 173-182). New York, NY: ACM.

Mimmack, G. M., Mason, S.J., Galpin, J.S. (2001). Choice of Distance Matrices in Cluster Analysis: Defining Regions. *Journal of Climate, 14*(12), 2790-2797.

Minelli, S., Del Secco, I., Naldi, G. . (2006). User Requirements Analysis - D 1.2 (Multimatch project). Retrieved from http://www.multimatch.org/docs/publicdels/D1.2Final.pdf

Minelli, S. H., Marlow, J., Clough, P., Cigarran Recuero, J.M., Gonzalo, J., Oomen, J., Loschiavo, D. (2007). Gathering Requirements for Multilingual Search of Audiovisual Material in Cultural Heritage. In Sorensen, L., van Kranenburg, H. (Eds.), *Proceedings of Workshop on User*

Centricity-State of the Art (16th IST Mobile and Wireless Communications Summit). Budapest, Hungary. Retrieved from http://www.multimatch.org/docs/papers/minelli_gathering.pdf.

MINERVAplus. (2006). Final Plan for Using and Disseminating Knowledge and Raise Public Participation and Awareness Report on Inventories and Multilingualism Issues: Multilingualism and Thesaurus - D6 (MINVERVAplus project). Retrieved from http://mek.oszk.hu/minerva/survey/delir20060130.pdf

Montgomery, A. L., Li, S., Srinivasan, K., Liechty, J.C. (2004). Modeling Online Browsing and Path Analysis Using Clickstream Data. *Marketing Science, 23*(4), 579-595.

Montgomery, A. L., Faloutsos, C. (2001). Identifying Web Browsing Trends and Patterns. *Computer, 34*(7), 94-95.

Munk, M. D., M. (2011). Influence of Different Session Timeouts Thresholds on Results of Sequence Rule Analysis in Educational Data Mining. In Cherifi, H. Z., J. M.; El-Qawasmeh, E. (Eds.), *Digital Information and Communication Technology and Its Applications - International Conference (DICTAP 2011)* (Communications in Computer and Information Science, Vol. 166). Berlin, Heidelberg: Springer.

Nadjarbashi-Noghani, M., Ghorbani, A. A. (2004). Improving the Referrer-Based Web Log Session Reconstruction. *2nd Annual Conference on Communication Networks and Services Research (CNSR 2004)* (pp. 286 - 292). Fredericton, N.B., Canada: IEEE Press.

Nicholas, D., Huntington, P., Jamali, H. R., Tenopir, C. (2006). What Deep Log Analysis Tells us About the Impact of Big Deal, Case Study OhioLink. *Journal of Documentation, 62*(4), 482-508.

Nicholas, D., Huntington, P., Jamali, H.R., Dobrowolski, T. (2007). Characterising and Evaluating Information Seeking Behaviour in a Digital Environment: Spotlight on the 'Bouncer'. *Information Processing & Management, 43*(4), 1085-1102.

Nicholas, D., Hungtington, P., Jamali, H.R. (2008). User Diversity: As Demonstrated by Deep Log Analysis. *The Electronic Library, 26*(1), 21-38.

Nicholas, D. (2009). Employing Deep Log Analysis to Evaluate the Information-Seeking Behaviour of Users of Digital Libraries. In Tsakonas, G., Papatheodorou, C, (Eds.), *Evaluation of Digital Libraries: An Insight to Useful Applications and Methods* (pp. 121-146). Oxford: Chandos.

Nicholas, D., Hungtington, P. (2010). Chapter 13. Evaluating the Use and Users of Digital Journal Libraries. In Papy, F. (Ed.), *Digital Libraries*. London, United Kingdom: ISTE Ltd.

Nicholas, D., Huntington, P., Jamali, H. R., & Watkinson, A. (2006). The Information Seeking Behaviour of the Users of Digital Scholarly Journals. *Information Processing & Management, 42*(5), 1345-1365.

Nicholson, S. (2004). A Conceptual Framework for the Holistic Measurement and Cumulative Evaluation of Library Services. *The Journal of Documentation, 60*(2), 164–182.

Oakes, M., Xu, Y. (2009). A Search Engine based on Query Logs and Search Log Analysis at the University of Sunderland. In Peters, C. (Ed.), *Results of the CLEF 2009 Cross-Language System Evaluation Campaign.* Corfu, Greece. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/85150/CLEF2009wn-LogCLEF-OakesEt2009.pdf.

Oard, D. W. (1997). Serving Users in Many Languages: Cross-Language Information Retrieval for Digital

Libraries. *D-Lib Magazine*. Retrieved from http://www.dlib.org/dlib/december97/oard/12oard.html

Oard, D. W., Diekema, A. (1998). Cross-Language Information Retrieval In Williams, M. (Ed.), *Annual Review of Information Science* (Vol. 33). Medford, NJ: Information Today Inc.

Oard, D. W., Gonzalo, J., Sanderson, M., López-Ostenero, F., Wang, J. (2004). Interactive Cross-Language Document Selection. *Information Retrieval, 7*(1-2), 205-228.

Oard, D. W. (2009). Multilingual Information Access In Bates, M. J., Maack, M.N. (Eds.), *Encyclopedia of Library and Information Sciences (3rd Ed.)*. Retrieved from http://terpconnect.umd.edu/~oard/pdf/elis09.pdf.

Ozmutlu, H. C., Spink, A., Ozmutlu, S. (2002). Analysis of Large Data Logs: An Application of Poisson Sampling on Excite Web Queries. *Information Processing & Management, 38*(4), 473-490.

Ozmutlu, S., Ozmutlu, H.C., Spink, A. (2009). From Analysis to Estimation of User Behavior. In Jansen, B. J., Spink, A., Taksai, I. (Eds.), *Handbook of Research on Web Log Analysis* (pp. 206-226). Hershey, USA: Information Science Reference.

Paolillo, J., Pimienta, D., Prado, D. (2007). Measuring linguistic diversity on the Internet. *UNESCO Publications for the World Summit on the Information Society*. Retrieved from http://unesdoc.unesco.org/images/0014/001421/142186e.pdf

Peinado, V., Artiles, J., Gonzalo, J., Barker, E., López-Ostenero, F. (2008). FlickLing: A Multilingual Search Interface for Flickr. In Peters, C. (Eds.), *Results of the CLEF 2008 Cross-Language System Evaluation Campaign*. Aarhus, Denmark. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86371/CLEF2008wn-iCLEF-PeinadoEt2008a.pdf.

Peters, C., Picchi, E. (1997). Across Languages, Across Cultures: Issues in Multilinguality and Digital Libraries. *D-Lib Magazine, 3*(5). Retrieved from http://www.dlib.org/dlib/may97/peters/05peters.html

Peters, C., Braschler, M., Clough, P.D. (2012). *Multilingual Information Retrieval: From Research to Practice*. Berlin, Heidelberg: Springer.

Peters, T. A. (1993). The History and Development of Transaction Log Analysis. *Library Hi Tech, 11*(2), 41-66.

Petras, V. (2011). Report on Multilingual Access Strategies to Digital Libraries - D 2.7.1 (EuropeanaConnect project). Retrieved from http://www.europeanaconnect.eu/documents/D2.7.1_eConnect-Facilitation%20and%20exchange%20of%20multilingual%20access%20strategies%20to%20digital%20libraries_v1.0.pdf

Petras, V., Ferro, N., Gäde, M., Isaac, A., Kleineberg, M., Masiero, I., Nicchio, M., Stiller, J. (2012). Cultural Heritage in CLEF (CHiC) Overview 2012. In Forner; P., K. J., Womser-Hacker, C. (Eds.), *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes*. Rome, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/155385/CLEF2012wn-CHiC-PetrasEt2012.pdf.

Petras, V., Bogers, T., Ferro, N., Masiero, I. (2013). Cultural Heritage in CLEF (CHiC) 2013 – Multilingual Task Overview. In Forner, P., Navigli, R., Tufis, D. (Eds*.), CLEF 2013 Evaluation*

Labs and Workshop, Online Working Notes. Valencia, Spain. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/430938/CLEF2013wn-CHiC-PetrasEt2013.pdf.

Petrelli, D., Beaulieu, M., Sanderson, M. (2002). User Requirement Elicitation for Cross-Language Information Retrieval. *The New Review of Information Behaviour Research, 2*, 17-35.

Petrelli, D. (2008). On the Role of User-Centred Evaluation in the Advancement of Interactive Information Retrieval. *Information Processing & Management, 44*(1), 22-38.

Purday, J. (2009). Think Culture: Europeana.eu from Concept to Construction. *The Electronic Library, 27*(6), 919–937.

Qiu, F., Liu, Z., Cho. J. (2005). Analysis of User Web Traffic with a Focus on Search Activities. In Doan, A., Neven, F., McCann, R., Bex, G.J. (Eds.), *8th International Workshop on the Web and Databases (WebDB)* (pp. 103 - 108). New York, NY: ACM.

Rau, P. L. P., Choongb, Y.-Y. Salvendyc, G. (2004). A Cross Cultural Study on Knowledge Representation and Structure in Human Computer Interfaces. *International Journal of Industrial Ergonomics, 34*(2), 117-129.

Renard, P. Y. (2007). ISO 2789 and ISO 11620: Short Presentation of Standards as Reference Documents in an Assessment Process. *Liber Quarterly, 17*(3/4). Retrieved from http://liber.library.uu.nl/index.php/lq/article/view/URN%3ANBN%3ANL%3AUI%3A10-1-113489/8101

Resnick, M. Vaughan, M. (2006). Best Practices and Future Visions for Search Interfaces. *Journal of the American Society for Information Science and Technology, 57*(6), 781-787.

Rose, D. E., Levinson, D. (2004). Understanding User Goals in Web Search. In Feldman, S., Uretsky, M., Najork, M., Wills, C. (Eds.), *Proceedings of the 13th International Conference on World Wide Web* (pp. 13-19). New York, NY, USA: ACM.

Sambandam, R. (2003). Cluster Analysis gets Complicated. *Marketing Research, 15*(1). Retrieved from http://www.trchome.com/component/content/article/66-published-articles/146-cluster-analysis.html

Saracevic, T. (2000). Digital Library Evaluation: Toward an Evolution on Concepts. *Library Trends, 49*(2), 350–369.

Saracevic, T. (2004). Evaluation of Digital Libraries: An Overview. *DELOS WP7 Workshop on the Evaluation of Digital Libraries*. Retrieved from http://comminfo.rutgers.edu/~tefko/DL_evaluation_Delos.pdf

Saulnier, A., Viaud, M.L. (2009). Evaluation Report of the Usability of the Europeana Web Site. Retrieved from http://pro.europeana.eu/c/document_library/get_file?uuid=ae1d74de-29c1-463c-887e-a6bc6ee0ed7a&groupId=10602

Shneiderman, B., Plaisant, C. (2010). *Designing the User Interface: Strategies for Effective Human-Computer Interaction (5th Ed.)*. Boston: Addison-Wesley.

Silverstein, C., Marais, H., Henzinger, M., Moricz, M. . (1999). Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum, 33*(1), 6-12.

Silvestri, F. (2010). Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval, 4*(1-2), 1-174.

Smith, G. (2008). *Tagging: people-powered metadata for the social web*. Berkeley, CA: New Riders.

Soergel, D. (1997). Multilingual Thesauri in Cross-Language Text and Speech Retrieval. In *Working Notes of AAAI Spring Symposium on Cross-Language Text and Speech Retrieval. Stanford*, CA. Retrieved from http://www.dsoergel.com/cv/B60.pdf.

Spiliopoulou, M., Mobasher, B., Berendt, B., Nakagawa, M. (2003). A Framework for the Evaluation of Session Reconstruction Heuristics in Web-Usage Analysis. *INFORMS Journal on Computing, 15*(2), 171-190.

Spink, A., Ozmutlu, S., Ozmutlu, H.C., Jansen, B.J. (2002). U.S. versus European Web Searching Trends. *SIGIR Forum, 36*(2), 32-38.

Spink, A., Jansen, B., & Ozmultu, C. (2000). Use of Query Reformulation and Relevance Feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy, 10*(4), 317-328.

Srinivasarao, V. (2008). Mining the Behavior of Users in a Multilingual Information Access Task. In Peters, C. (Eds.), *Results of the CLEF 2008 Cross-Language System Evaluation Campaign*. Aarhus, Denmark. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86371/CLEF2008wn-iCLEF-Vundavalli2008.pdf.

Stassopoulou, A., Dikaiakos, M.D. . (2007). A Probabilistic Reasoning Approach for Discovering Web Crawler Sessions. In Dong, G., Lin, X., Wang, W., Yang, Y., Yu, J.X. (Eds.), *Advances in Data and Web Management. Proceedings of the Joint 9th Asia-Pacific Web Conference (APWeb 2007) and 8th International Conference, on Web-Age Information Management (WAIM 2007)* (Lecture Notes in Computer Science, Vol. 4505, pp. 265-272). Berlin, Heidelberg: Springer.

Stiller, J., Gaede, M., Petras, V. (2010). Ambiguity of Queries and the Challenges for Query Language Detection. In Braschler, M., Harman, D., Pianta, E. (Eds.), *CLEF 2010 Labs and Workshops Notebook Papers*. Padua, Italy. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86374/CLEF2010wn-LogCLEF-StillerEt2010.pdf.

Stiller, J., Gäde, M., Petras, V. (2011). Is Tagging Multilingual? A Case Study with BibSonomy. In Newton, G., Wright, M., Cassel, L. (Eds.), *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries* (pp. 421-422). New York, NY: ACM.

Stiller, J. (2012). A Framework for Classifying Interactions in Cultural Heritage Infomration Systems. *International Journal of Heritage in the Digital Era: Proceedings of EUROMED 2012: Progress in Cultural Heritage Preservation,* (Supplement 1), 141-146.

Stiller, J., Gäde, M., Petras, V. (2013). Multilingual Access to Digital Libraries: The Europeana Use Case. *Information - Wissenschaft & Praxis, 64*(2-3), 86 - 95.

Stiller, J., Gäde, M. (to be published). From Eiffelturm to Tour Eiffel: Query Reformulation in a Multilingual Digital Library.

Strohmaier, M., Kröll, M. (2012). Acquiring Knowledge about Human Goals from Search Query Logs. *Information Processing & Management, 48*(1), 63-82.

Taksa, I., Spink, A., Jansen, B.J. . (2009). Web Log Analysis: Diversity of Research Methodologies. In Taksa, I., Spink, A., Jansen, B.J. (Eds.), *Handbook of Research on Web Log Analysis* (pp. 506-522). Hershey, USA: Information Science Reference.

Tan, P. N., Kumar, V. (2002). Discovery of Web Robot Sessions Based on their Navigational Patterns.

Data Mining and Knowledge Discovery, 6(1), 9-35.

Tanase, D. I., Kapetanios, E. . (2008). Evaluating the Impact of Personal Dictionaries for Cross-Language Information Retrieval of Socially Annotated Images. In Peter, C. (Eds.), *Results of the CLEF 2008 Cross-Language System Evaluation Campaign*. Aarhus, Denmark. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/86371/CLEF2008wn-iCLEF-TanaseEt2008.pdf.

Tenopir, C. (2003). Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies. Retrieved from http://www.clir.org/pubs/reports/pub120/pub120.pdf

Toms, E. G., Hall, M.M. (2013). The CHiC Interactive Task (CHiCi) at CLEF2013. In Forner, P., Navigli, R., Tufis, D. (Eds.), CLEF 2013 Evaluation Labs and Workshop, Online Working Notes. Valencia, Spain. Retrieved from http://ims-sites.dei.unipd.it/documents/71612/430938/CLEF2013wn-CHiC-TomsEt2013.pdf.

Trojahn, C., Siciliano, L. (2009). User Requirements for Advanced Features - D7.4 (CACAO project). Retrieved from http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D7.4.pdf

Tsakonas, G., Kapidakis, S., Papatheodorou, C. (2004). Evaluation of User Interaction in Digital Libraries In Agosti, M., Fuhr, N. (Eds.), *Notes of the DELOS WP7 Workshop on the Evaluation of Digital Libraries* (pp. 45 - 60). Padova, Italy Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.101.9642&rep=rep1&type=pdf

Tsakonas, G., Papatheodorou, C. (Ed.). (2009). *Evaluation of Digital Libraries: An Insight into Useful Aplications and Methods*. Great Abington, UK: Chandos.

UNESCO (2003a). Charter on the Preservation of Digital Heritage Records of the General Conference - 32 C/Resolution 15, Volume 1 (pp. 74-77). Paris. Retrieved from http://unesdoc.unesco.org/images/0013/001331/133171e.pdf#page=80.

UNESCO (2003b). Recommendations Concerning the Promotion and Use of Multilingualism and Universal Access in Cyberspace Records of the General Conference - 32 C/Resolution 15, Volume 1 (pp. 70-74). Paris Retrieved from http://unesdoc.unesco.org/images/0013/001331/133171e.pdf#page=80.

Vassilakaki, E., Johnson, F., Hartley, R.J., Randall, D. (2009). Users' Perceptions of Searching in Flickling. In Peters, C. (Eds.), *Results of the CLEF 2009 Cross-Language System Evaluation Campaign.* Corfu, Greece. Retrieved from http://clef.isti.cnr.it/2009/working_notes/vassilaki.pdf.

Vassilakaki, E., Garoufallou, E. (2013). Multilingual Digital Libraries: A Review of Issues in System-Centered and User-Centered Studies, Information Retrieval and User Behavior. *The International Information & Library Review, 45*(1-2), 3-19.

Voorbij, H. (2010). The Use of Web Statistics in Cultural Heritage Institutions. *Performance Measurement and Metrics, 11*(3), 266-279.

Waller, V. (2009). What Do the Public Search for on the Catalogue of the State Library of Victoria? *Australian Academic & Research Libraries, 40*(4), 266-285.

Wang, P. (1999). Methodologies and Methods for User Behavioral Research. *Annual Review of Information Science and Technology (ARIST), 34*(1), 53-99.

Ward, J. H. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association, 58*(301), 236–244.

White, R. W., Drucker, S.M. (2007). Investigating Behavioral Variability in Web Search. In Fridman Noy, N., Alani, H., Stumme, G., Mika, P., Sure, Y., Vrandecic, D (Eds.), *Proccedings of the 16th International Conference in World Wide Web* (pp. 21-30). New York, NY: ACM.

White, A., Kamal, E.D. (2006). *E-Metrics for Library and Information Professionals: How to Use Data for Managing and Evaluating Electronic Resource Collections*. New York: Facet Publishing.

Wilson, M., L. (2011). Search User Interface Design. *Synthesis Lectures on Information Concepts, Retrieval, and Services, 3*(3), 1-143.

Wright, S. P. (1992). Adjusted P-Values for Simultaneous Inference. *Biometrics, 48*, 1005 - 1013.

Wu, D., He, D., Lou, B. (2012). Multilingual Needs and Expectations in Digital Libraries: A Survey of Academic Users with Different Languages. *The Electronic Library, 30*(2), 182-197.

Xie, H. (2006). Evaluation of Digital Libraries: Criteria and Problems from Users' Perspectives. *Library and information Science Research, 28*(3), 433-452.

Xie, H. (2008). Users' Evaluation of Digital Libraries: Their Uses, Their Criteria, and Their Assessment. *Information Processing & Management, 44*(3), 1346-1373.

Zahedi, F., Van Pelt, W., Song, J. (2001). A Conceptual Framework for International Web Design. *IEEE Transactions on Professional Communication 44*(2), 83-103.

Zhang, J., Ghorabi, A. A. (2004). The Reconstruction of User Sessions from a Server Log using Improved Time-Oriented Heuristics. In Ghorbani, A. A., Lewis, G. (Eds.), *Proceedings of the 2nd Annual Conference on Communication Networks and Services Research (CNSR 2004)* (pp. 315 - 322). Fredericton, Canada: IEEE Computer Society.

Zhang, J., Lin, S. (2007). Multiple Language Supports in Search Engines. *Online Information Review, 31*(4), 516-532.

Zoe, L. R., Dimartino, D. (2000). Cultural Diversity and End-User Searching: An Analysis of Gender and Language Background. *Research Strategies 17*(4), 291-305.

APPENDICES

A. COUNTRY PROFILES

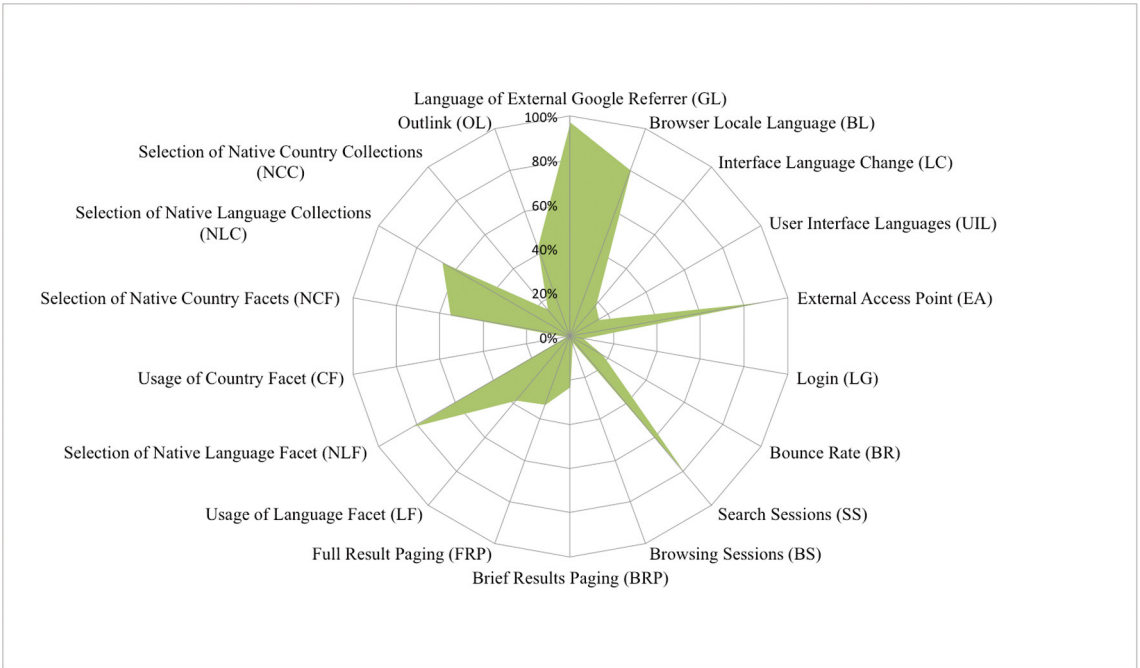

Figure A 1 Austria country profile

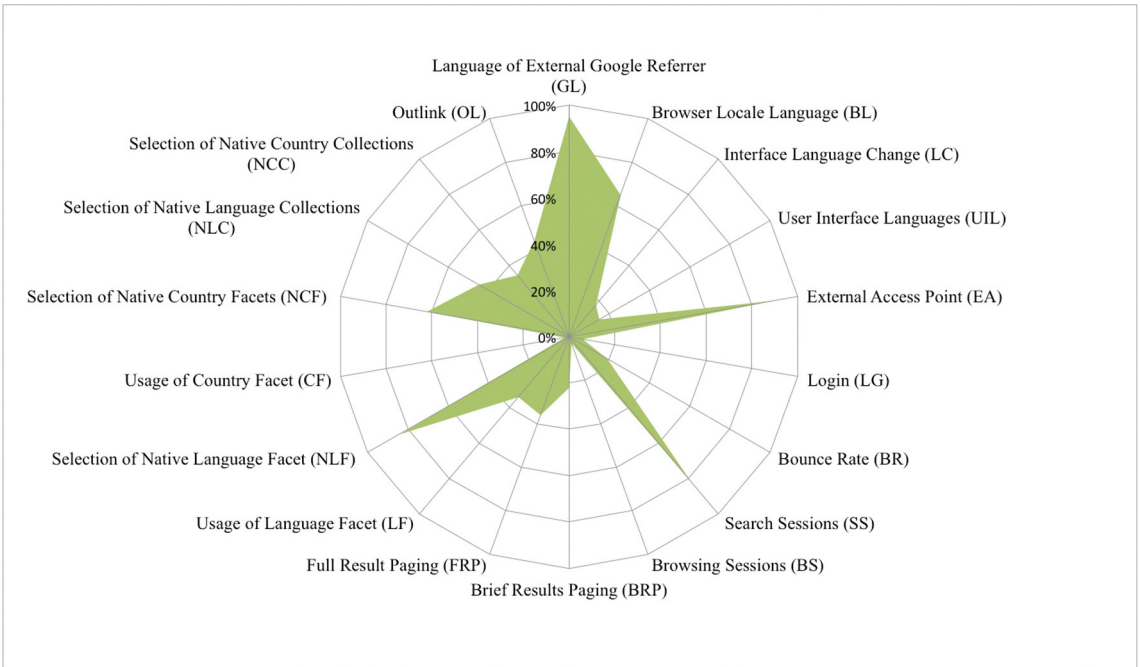

Figure A 2 Belgium country profile

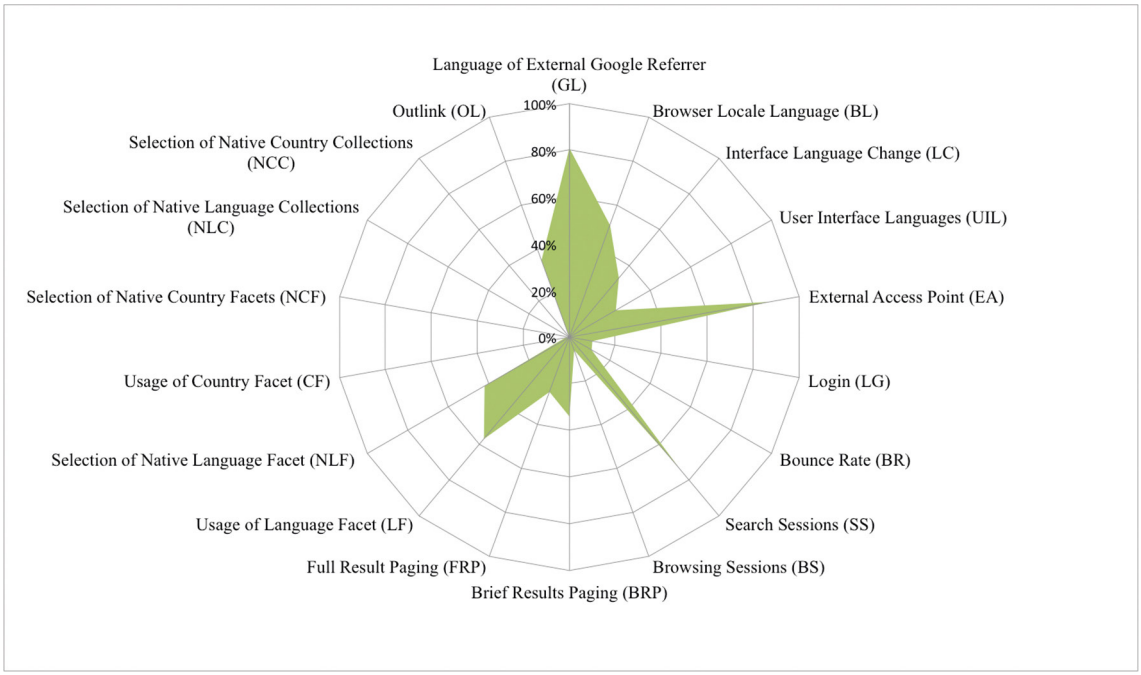

Figure A 3 Brazil country profile

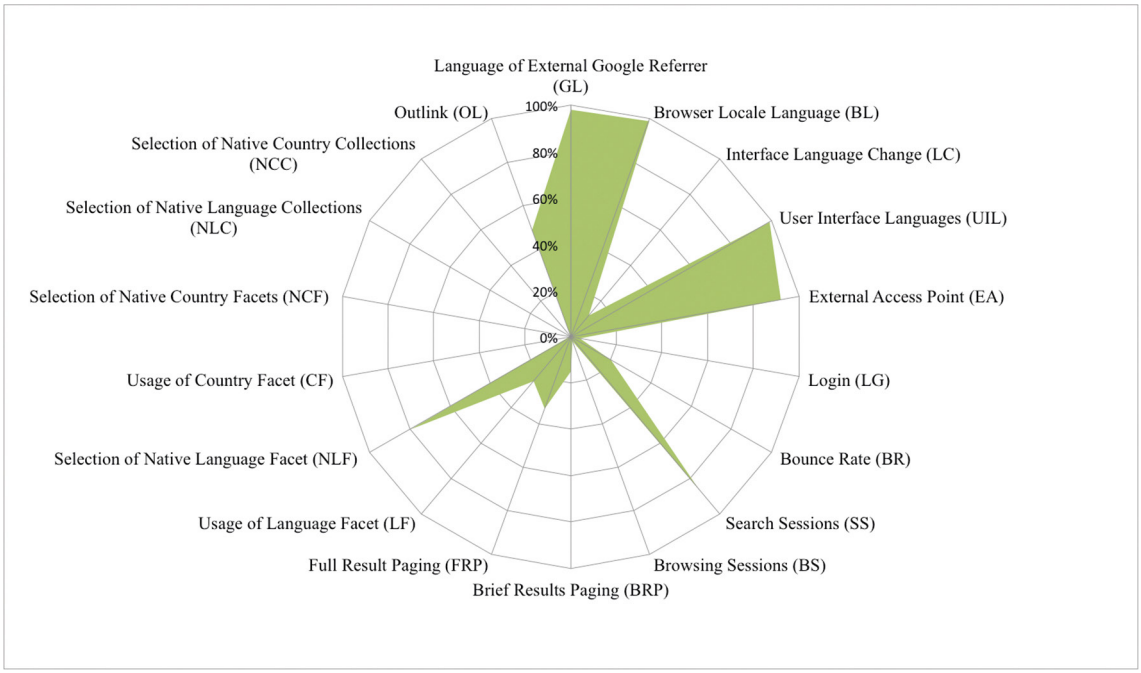

Figure A 4 Canada country profile

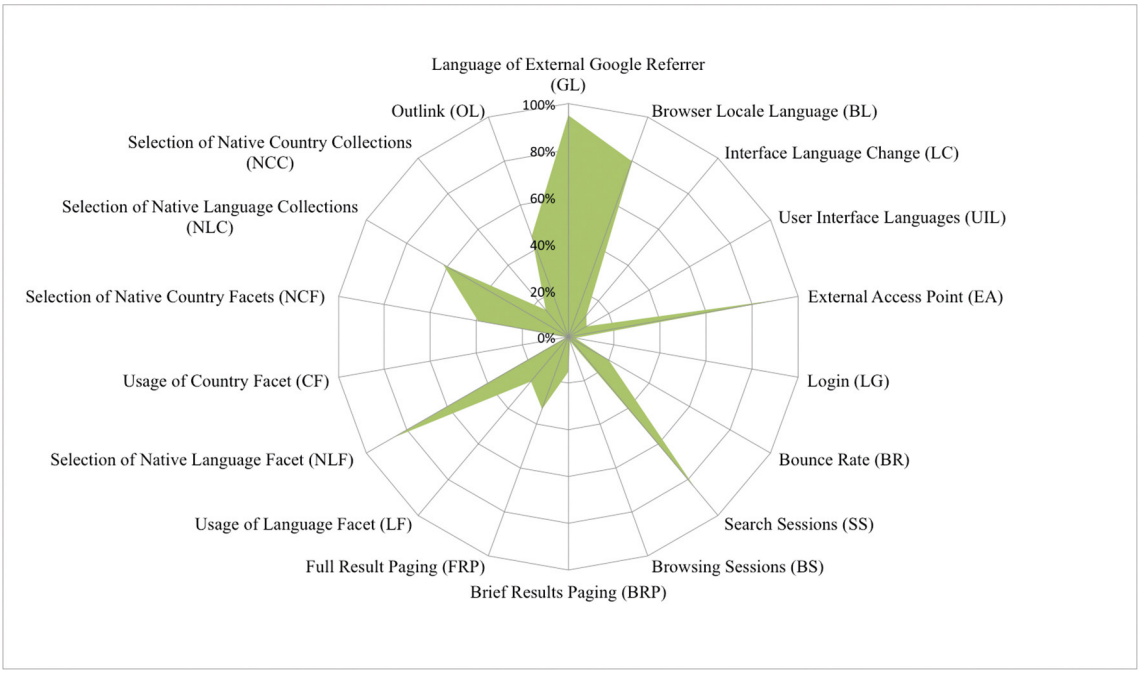

Figure A 5 Switzerland country profile

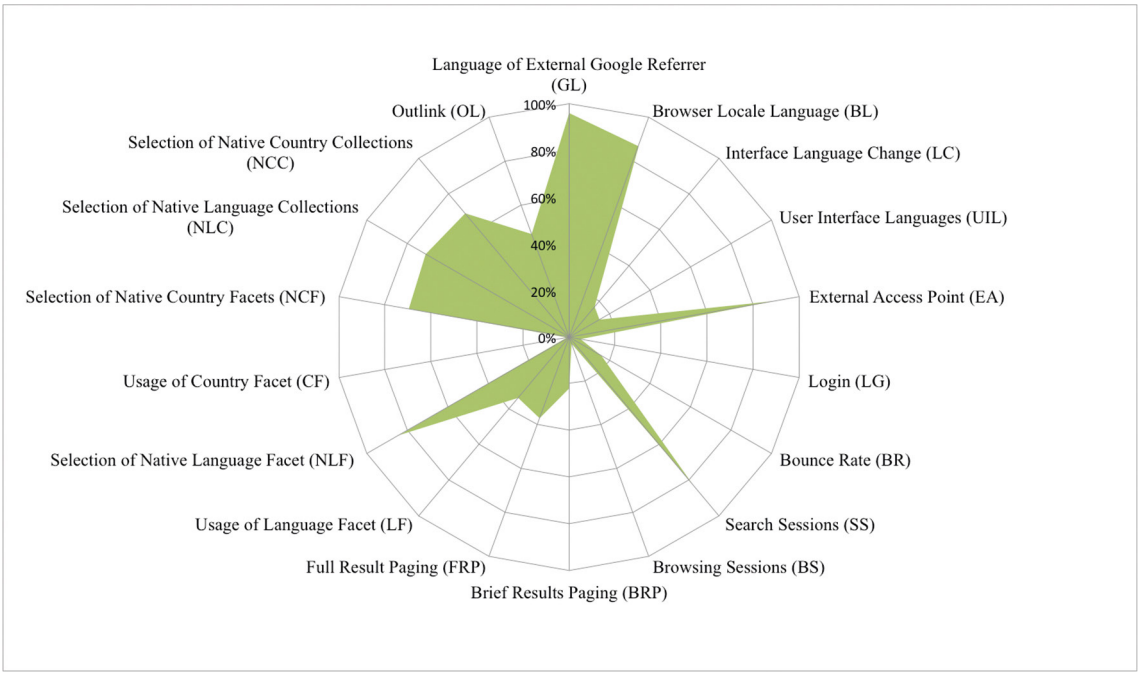

Figure A 6 Germany country profile

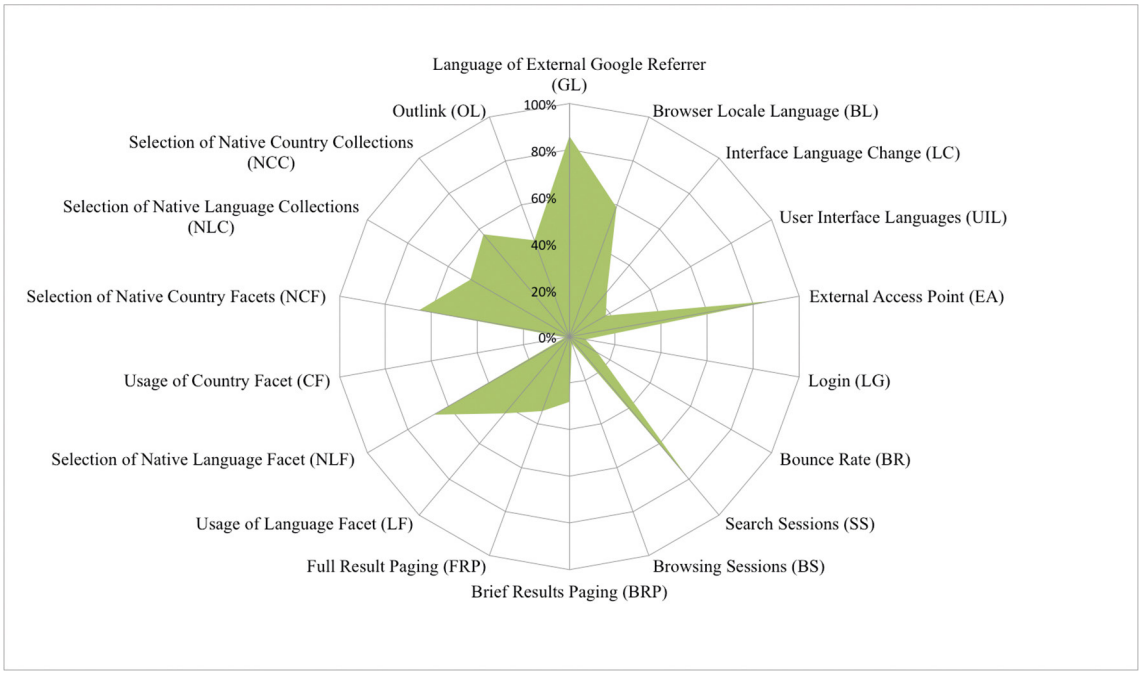

Figure A 7 Spain country profile

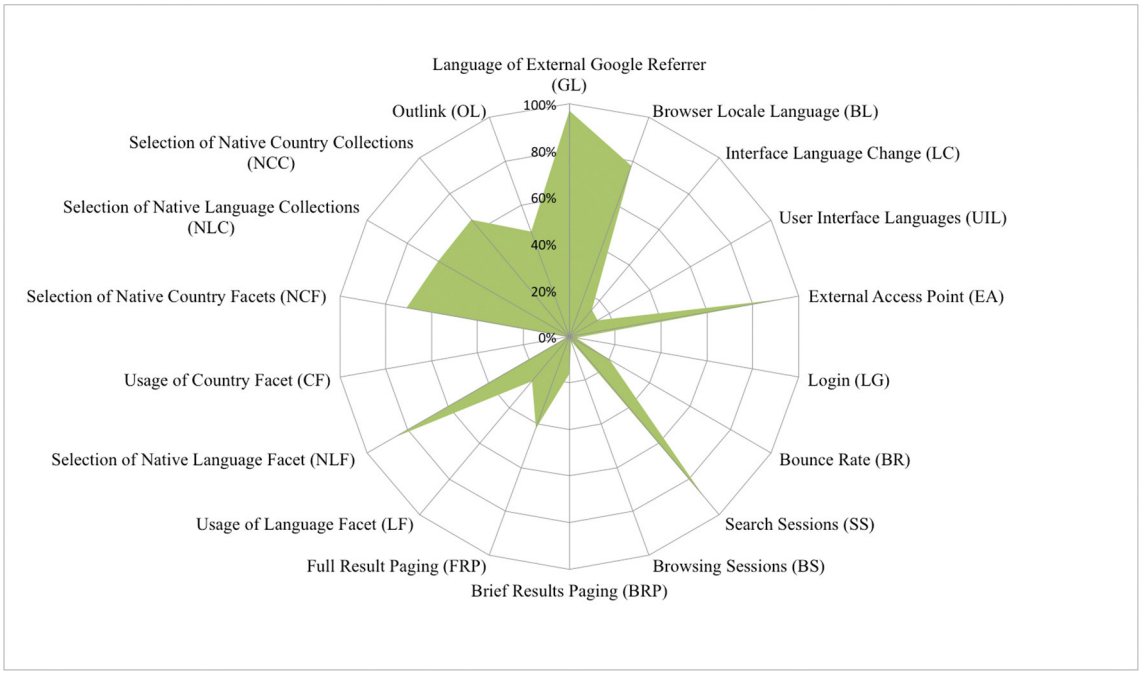

Figure A 8 France country profile

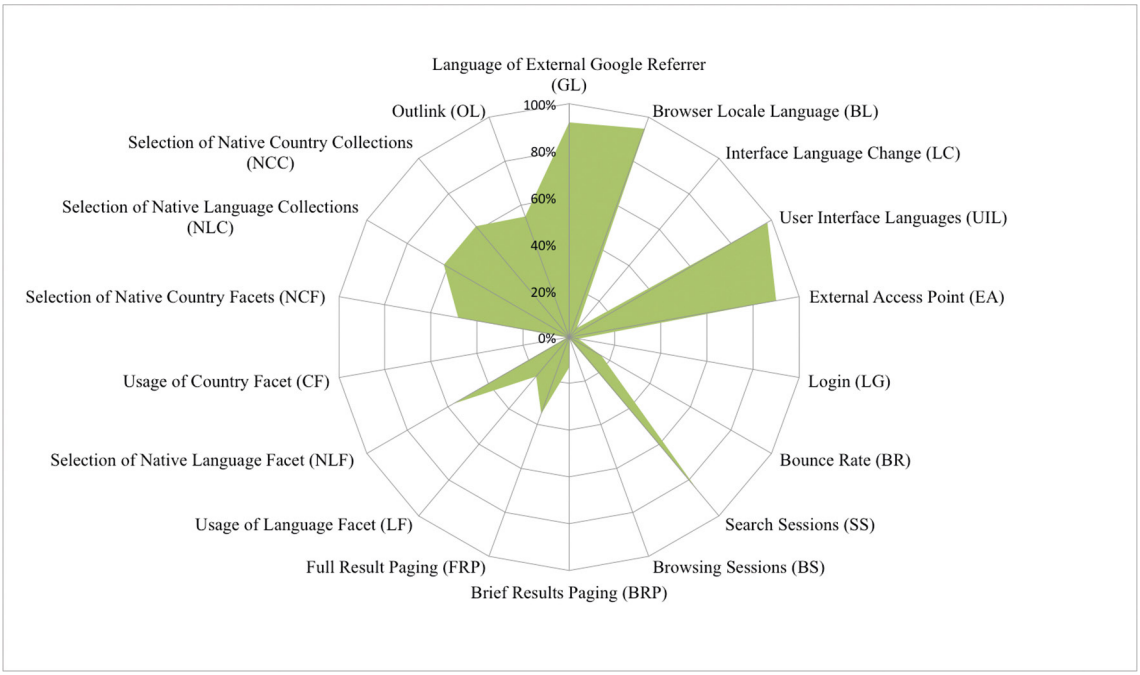

Figure A 9 Great Britain country profile

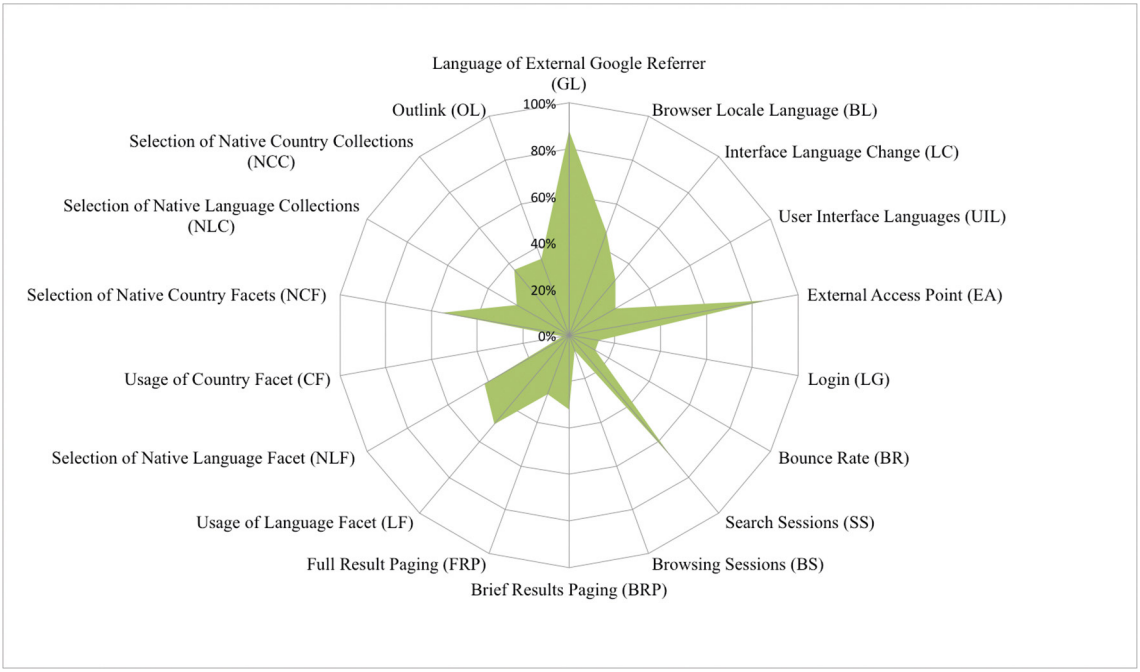

Figure A 10 Greece country profile

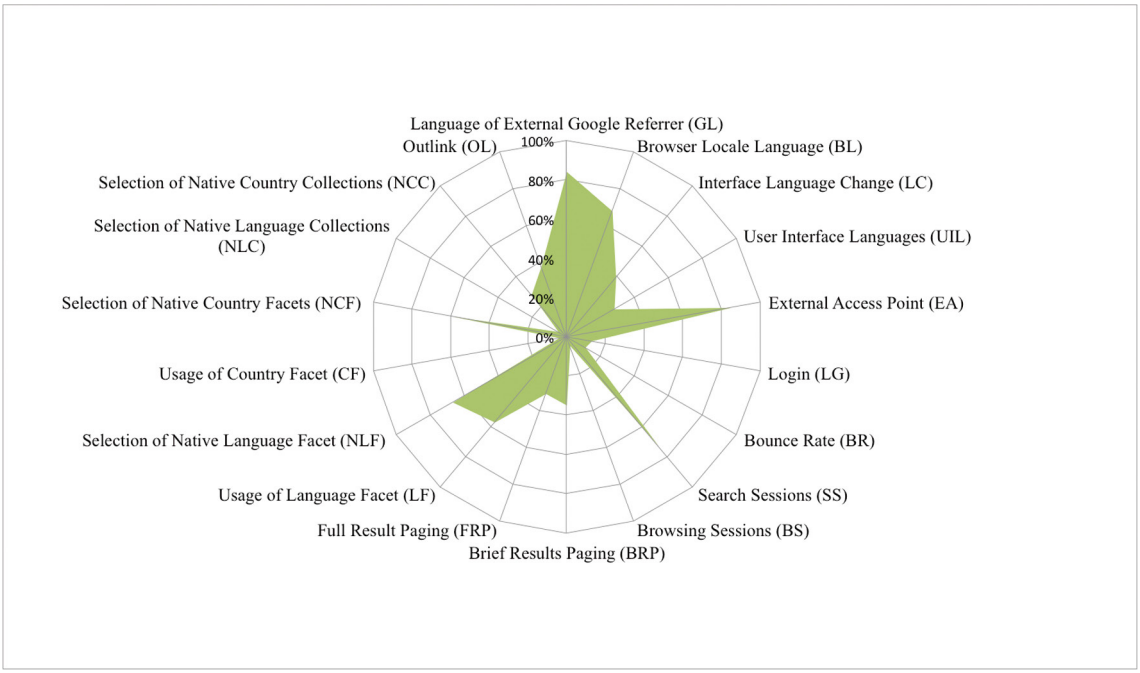

Figure A 11 Hungary country profile

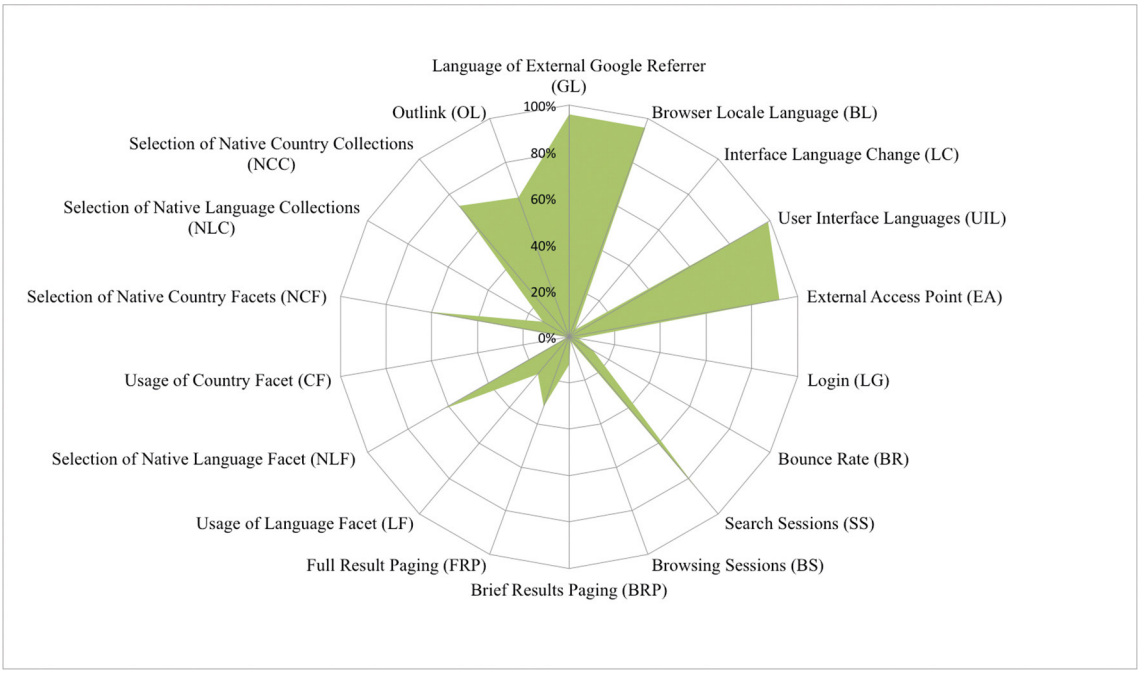

Figure A 12 Ireland country profile

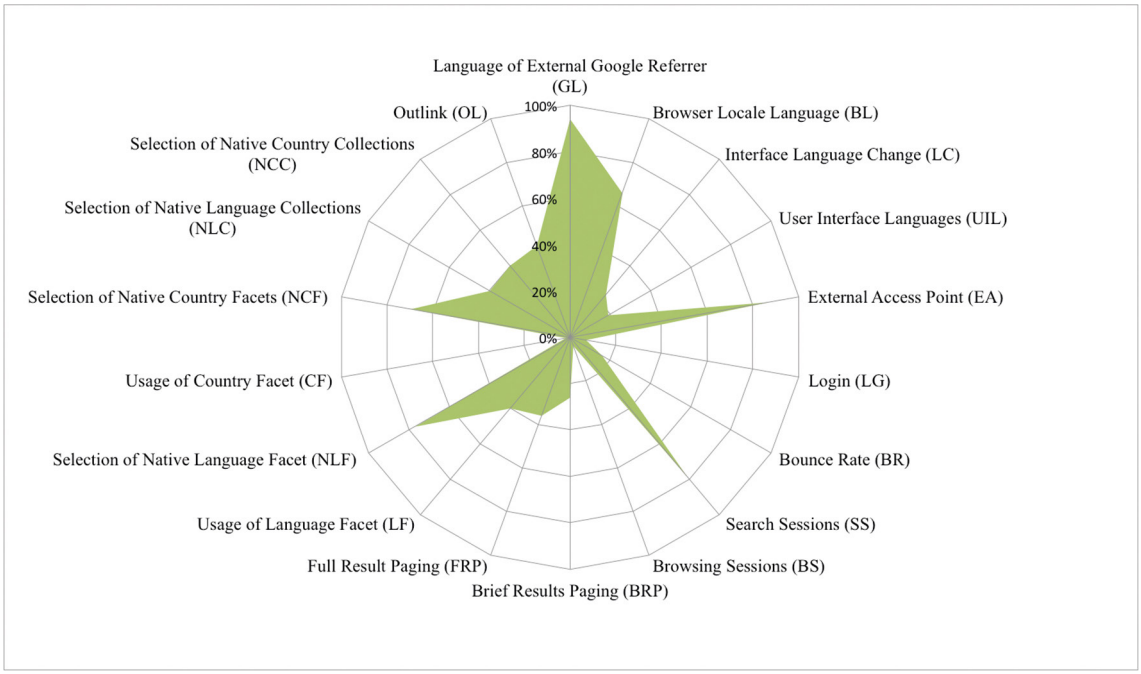

Figure A 13 Italy country profile

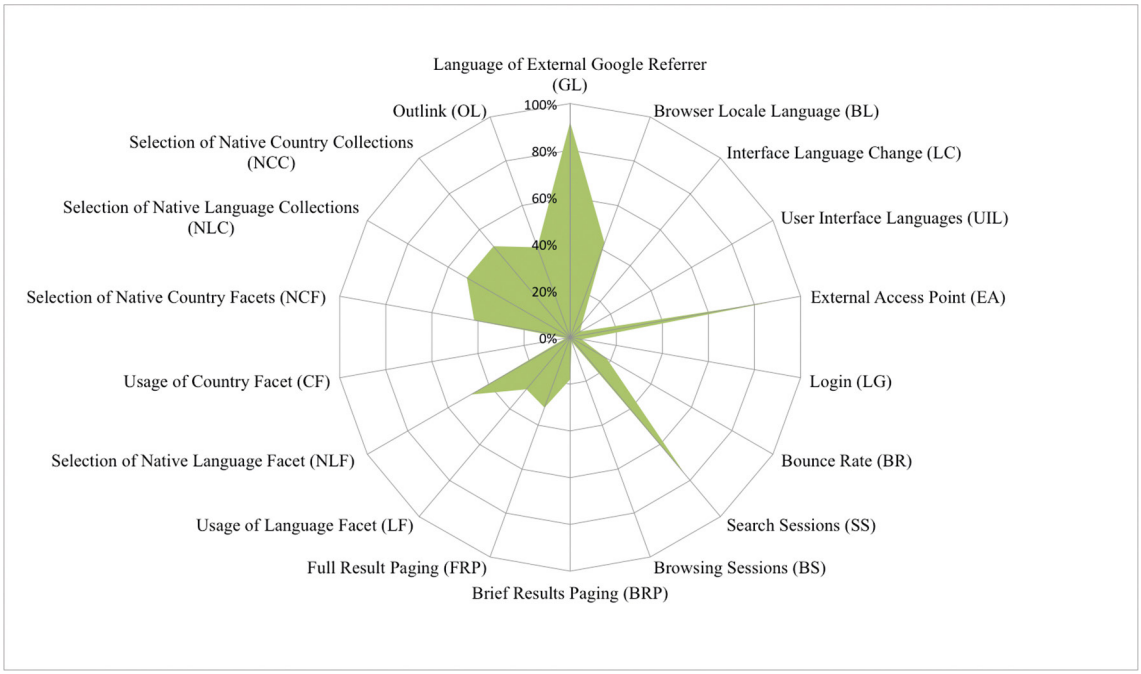

Figure A 14 Netherlands country profile

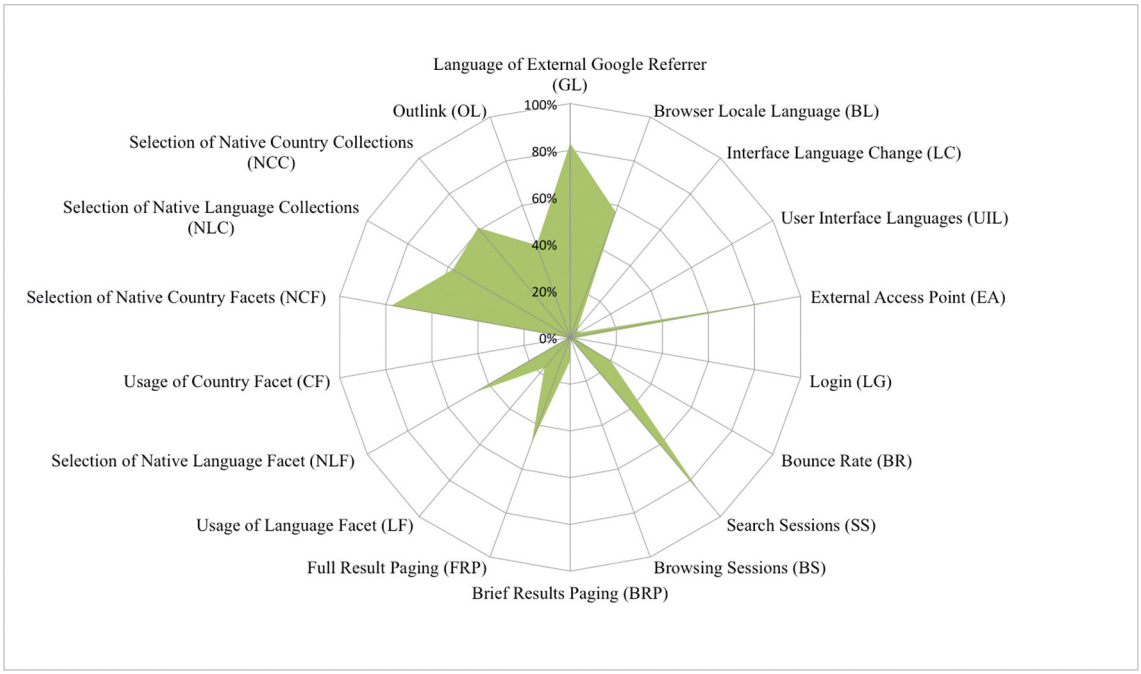

Figure A 15 Norway country profile

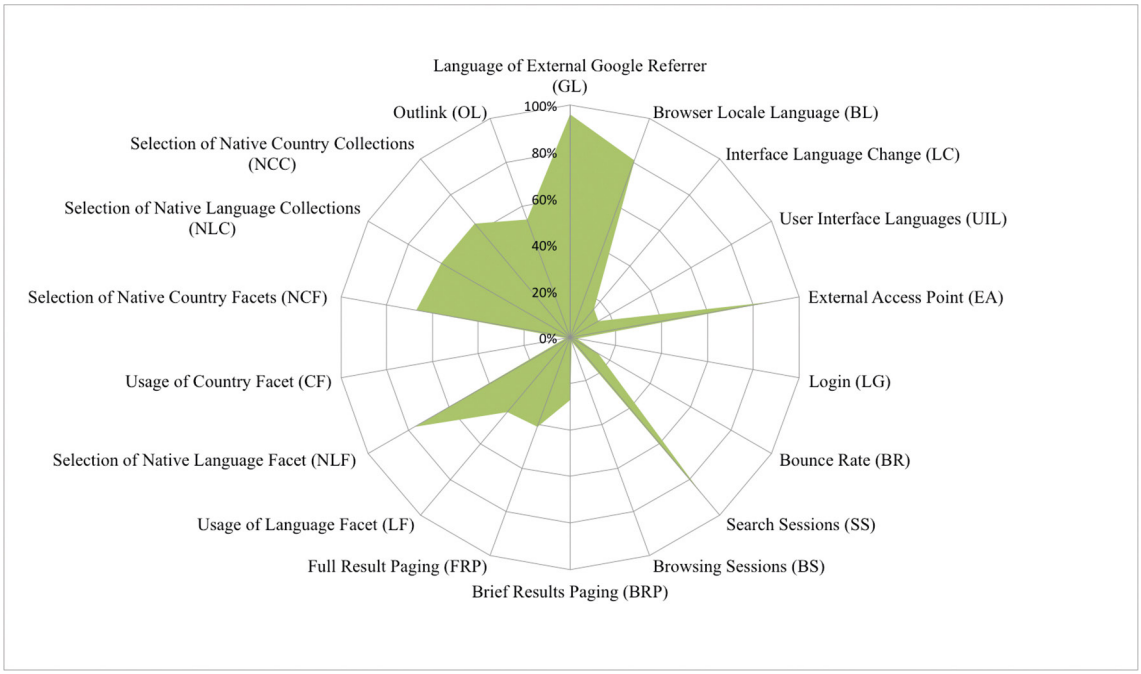

Figure A 16 Poland country profile

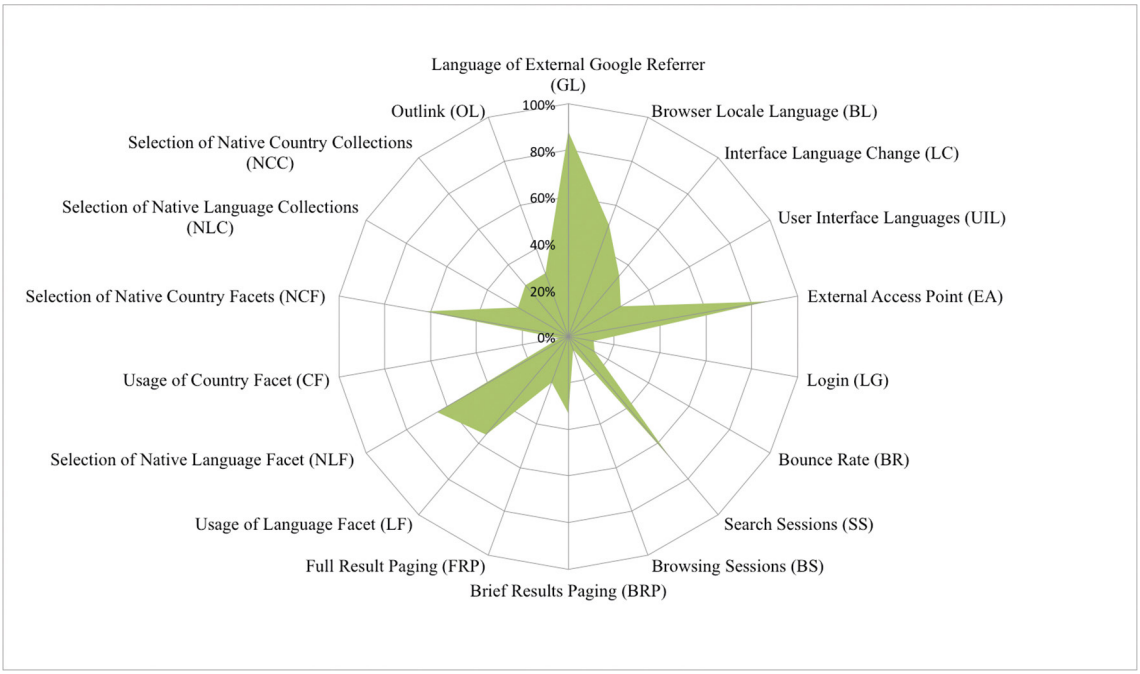

Figure A 17 Portugal country profile

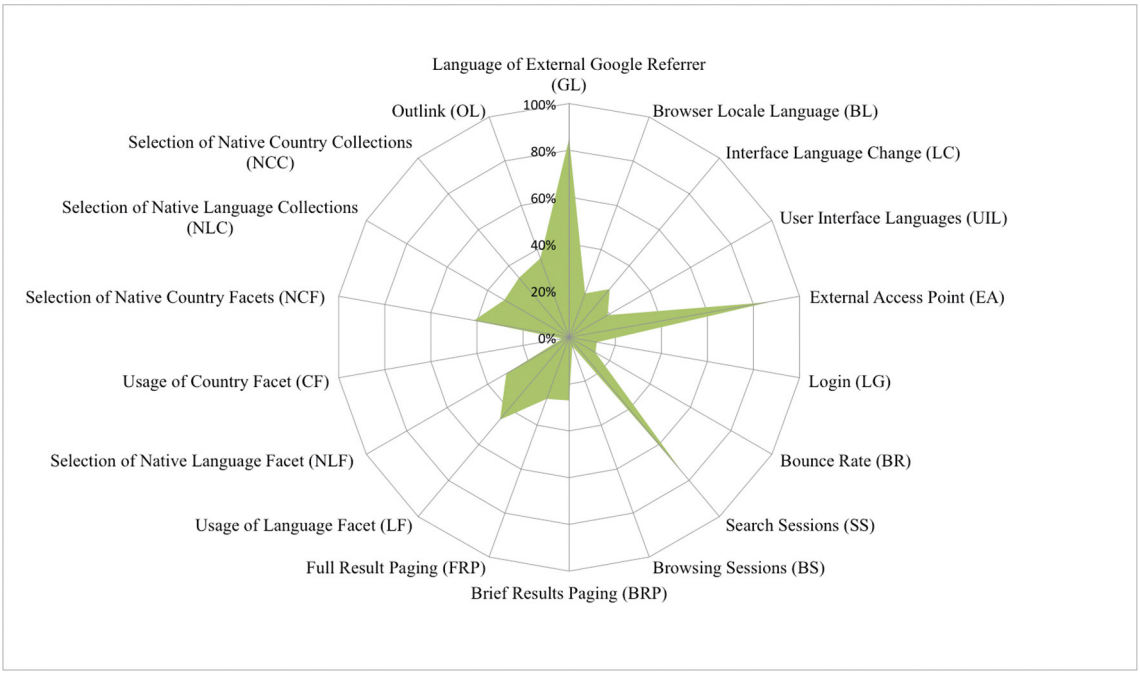

Figure A 18 Romania country profile

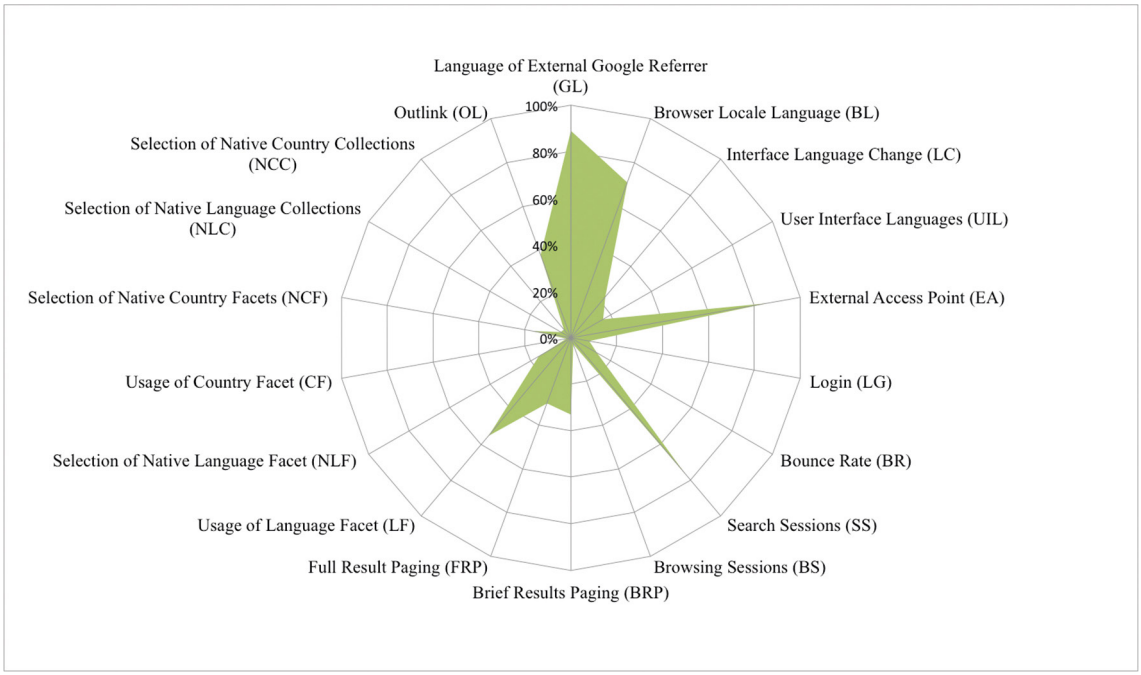

Figure A 19 Russia country profile

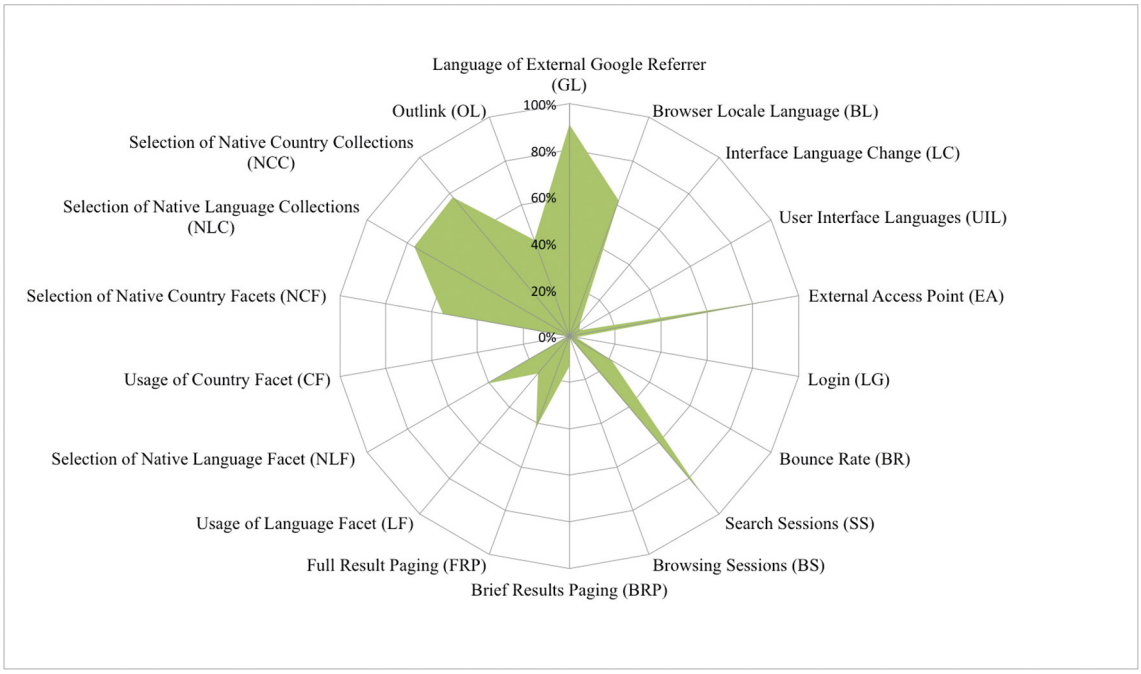

Figure A 20 Sweden country profile

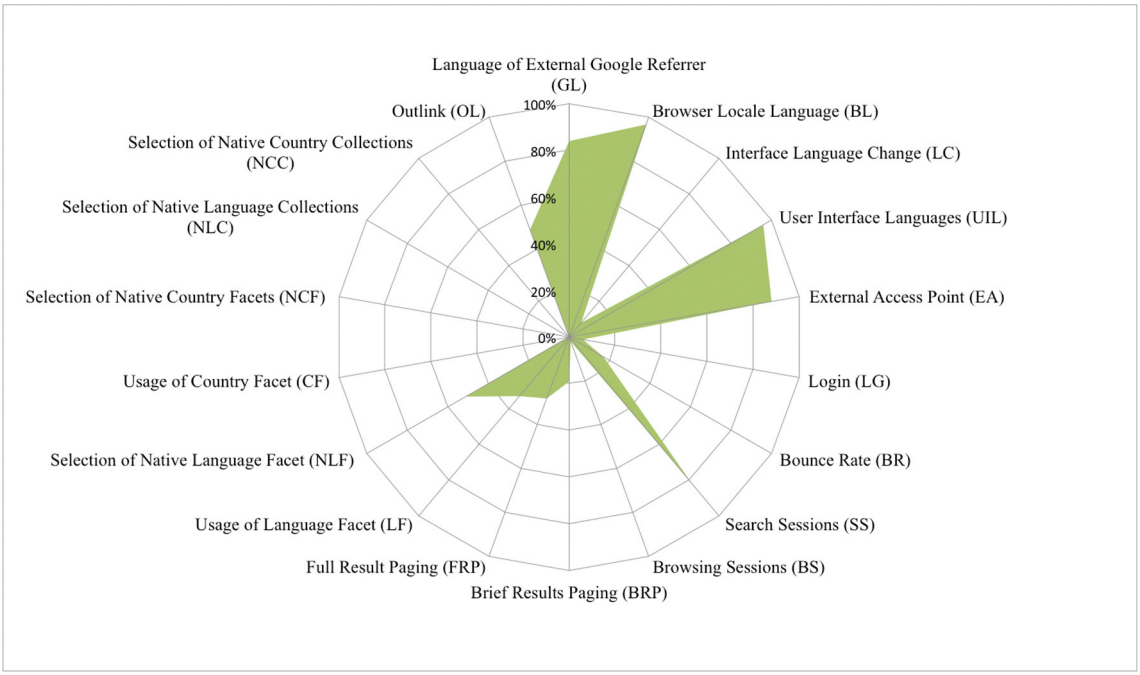

Figure A 21 US country profile

B. RESULTS FOR ALL VARIABLES PER COUNTRY

Table B.1 summarizes the results for all variables per country. The columns contain values per variable, each row presenting one country. Looking at the first row and first five columns, the results for multilingual interface variables for Austria are presented. The second column presents the percentage of native language Google accesses (GL) with 97% of all Google referrers, column three the native browser sessions (BL) (80%), column three the percentage of sessions with an interface language change (LC) (18%) and in column four it can be seen that 15% of all Austrian sessions are conducted with the native interface language (UIL).

	Multilingual Interface				Multilingual Search and Browsing									Multilingual Result Representation							
V	GL	BL	LC	UIL	D	Q	EA	BR	LG	SS	BS	BRP	FRP	LF	NLF	CF	NCF	NLC	NCC	OL
AT	97%	8%	18%	15%	23.73	1.82	89%	17%	6%	8%	3%	23%	33%	38%	82%	2%	55%	67%	16%	43%
BE	95%	65%	18%	5%	24.4	1.71	9%	19%	6%	82%	3%	22%	36%	34%	84%	2%	62%	45%	35%	44%
BR	81%	51%	33%	23%	25.59	1.89	87%	11%	1%	73%	6%	34%	25%	57%	42%	2%	x[63]	9%	x	35%
CA	98%	99%	12%	99%	17.71	1.44	92%	2%	4%	86%	2%	15%	33%	25%	80%	1%	x	34%	x	49%
CH	0.95	80%	12%	9%	26.00	1,51	92%	2%	3%	84%	2%	15%	33%	25%	86%	1%	39%	62%	15%	46%
DE	96%	87%	17%	15%	21.91	1.77	89%	16%	5%	82%	3%	22%	37%	34%	84%	1%	70%	71%	69%	47%
ES	86%	59%	25%	18%	23.43	1.94	88%	14%	7%	78%	3%	28%	34%	43%	67%	2%	65%	49%	57%	44%
FR	97%	78%	15%	14%	23.85	1.64	92%	20%	3%	90%	2%	16%	42%	25%	86%	1%	71%	65%	66%	48%
GB	92%	95%	5%	98%	17.81	1.49	90%	16%	4%	85%	1%	13%	35%	22%	57%	1%	48%	62%	62%	55%
GR	88%	0.47	31%	23%	27.27	1.90	86%	13%	13%	68%	7%	32%	27%	50%	42%	3%	56%	26%	37%	35%
HU	84%	68%	39%	28%	20.75	2.06	85%	11%	13%	73%	5%	0,35	31%	57%	67%	4%	60%	5%	28%	38%
IE	96%	96%	4%	99%	19.67	1.42	92%	12%	4%	86%	2%	0,12	32%	21%	62%	1%	63%	13%	74%	64%
IT	94%	66%	24%	19%	23,03	1.89	87%	16%	7%	79%	4%	26%	36%	40%	77%	2%	70%	40%	40%	42%

[63] For Brazil, Canada and the US no native country content was available within Europeana

	Multilingual Interface					Multilingual Search and Browsing									Multilingual Result Representation							
V	**GL**	**BL**	**LC**	**UIL**	**D**	**Q**	**EA**	**BR**	**LG**	**SS**	**BS**	**BRP**	**FRP**	**LF**	**NLF**	**CF**	**NCF**	**NLC**	**NCC**	**OL**
NL	92%	43%	8%	5%	30.6	1.76	88%	18%	5%	76%	2%	18%	32%	29%	49%	2%	42%	51%	51%	41%
NO	83%	57%	5%	4%	16.99	1.74	91%	20%	2%	86%	1%	10%	48%	17%	47%	1%	78%	58%	61%	42%
PL	96%	81%	16%	14%	26.28	2.18	88%	14%	3%	88%	2%	27%	41%	42%	77%	2%	67%	64%	64%	54%
PT	88%	51%	34%	26%	24.38	1.7	87%	13%	11%	68%	6%	33%	21%	55%	65%	4%	63%	25%	29%	29%
RO	85%	2%	27%	19%	19.7	1.84	0,87	13%	12%	75%	4%	27%	28%	46%	31%	3%	41%	32%	33%	36%
RU	89%	71%	23%	16%	38.92	2.14	86%	13%	8%	77%	3%	33%	30%	55%	16%	2%	17%	5%	5%	38%
SE	91%	62%	7%	5%	19,67	1.63	90%	21%	3%	88%	1%	13%	42%	21%	41%	1%	55%	77%	78%	44%
US	84%	97%	8%	96%	22.18	1.56	88%	17%	6%	81%	2%	19%	28%	33%	51%	1%	x	38%	x	49%
M	91%	69%	18%	31%	23.52	1.76	89%	6%	16%	80%	3%	23%	34%	37%	62%	2%	49%	43%	39%	44%

Table B 1 Summary of all variables per country (percentage of usage or selection of native language / country over all sessions, except for D (session duration in minutes) and Q (number of queries per sessions)) [64].

[64] V (Variable), GL (Language of External Google Referrer), BL (Browser Locale Language), LC (Interface Language Change), UIL (User Interface Language), D (Duration of Sessions), Q (Unique Queries per Session), EA (External Access Point); BR (Bounce Rate), LG (Login), SS (Search Sessions), BS (Browsing Sessions), BRP (Brief Result Paging), FRP (Full Result Paging), LF (Usage of Language Facet), NLF (Selection of Native Language Facet), CF (Usage of Country Facet), NCF (Selection of Native Country Facet), NLC (Selection of Native Language Collections), NCC (Selection of Native Country Collections), OL (Outlinks to Content provider)

C. EUROPEANA ACTIONS

Action	Explanation
"BRIEF_RESULT_FROM_PACTA"	Result list generated from the PACTA (people are currently thinking about) function.
"FULL_RESULT"	User clicks on an object from the result list and is directed to the detailed full result presentation.
"INDEXPAGE"	User views the homepage.
"REDIRECT_OUTLINK"	User clicks on a link directing to the content provider.
"RETURN_TO_RESULTS"	Users from a full view to the result list.
"STATICPAGE"	Pages views with no dynamic parameter such as “contact us”, “terms of use and policies”.
"BRIEF_RESULT"	Result list from a search query.
"LANGUAGE_CHANGE"	A user switches the interface language via drop down menu.
"LOGIN"	User logs in the user profile MyEuropeana.
"TIMELINE"	Users views objects at a timeline.
"LOGOUT"	User logs out from MyEuropeana.
"LOGOUT_COOKIE_THEFT"	Matching of cookies fails.
"REGISTER"	User registers for MyEuropeana.
"REGISTER_FAILURE"	Registration process failed.
"REGISTER_SUCCESS"	The registration was successfully confirmed.
"EXCEPTION_CAUGHT"	Requested object was not found.
"CONTACT_PAGE"	User clicks on the contact / feedback page.
"FEEDBACK_SEND"	User has send a message at the feedback page.
"FULL_RESULT_FROM_TIME_LINE_VIEW"	Object view presented at the timeline.
"MY_EUROPEANA"	MyEuropeana link for registration or log in.
"FEEDBACK_SEND_FAILURE"	User message from the feedback page.
"CHANGE_PASSWORD_SUCCES"	User changed the log in password.
"SITE_MAP_XML"	XML rendition of the sitemap so crawlers can navigate the website.
"SAVE_ITEM"	Users saves an object in MyEuropeana.

"SAVE_SOCIAL_TAG"	Users assigns a tag to an object saved in MyEuropeana.
"CHANGE_PASSWORD_FAILURE"	Change of user profile password failed.
"FULL_RESULT_FROM_SAVED_ITEM"	A user clicks on a saved object and views the full result representation.
"SAVE_SEARCH"	User saves previous search terms in MyEuropeana.
"BRIEF_RESULT_FROM_SAVED_SEARCH"	User clicks on a previous saved search and receives a result lis.t
"REMOVE_SAVED_ITEM"	Users removes a saved object from MyEuropeana.
"REMOVE_SOCIAL_TAG"	User removes a tag from MyEuropeana.
"FULL_RESULT_FROM_CAROUSEL"	Users clicks on an object from the carousel. and is directed to the full result presentation
"ERROR_TOKEN_EXPIRED"	User session is terminated due to inactivity.
"REMOVE_SAVED_SEARCH"	User removes saved search terms from MyEuropeana.
"AJAX_ERROR"	Error with Java Script request from front-end to back-end service.
"BROWSE_BOB"	Carrousel / timeline (replaced by timeline).
"YEAR_GRID"	Functionality before timeline.
"FULL_RESULT_FROM_YEAR_GRID"	User views an object presented at the year grid (now time line).
"REDIRECT_TO_SECURE"	User is directed to a secure communication page (HTTPS).
"BROWSE_ALL"	Users browses through the time line without selecting (a) specific year(s).
"FULL_RESULT_HMTL"	The HTML rendering of the full result page.
"SHOW_SIWA_MENU"	The SIWA menu allows you to enrich the full-view with external web services (only in test version).
"FULL_RESULT_EMBEDED"	Full result is called and embedded into an external page (only in test version).
"FULL_RESULT_SRW"	The XML SRU/SRW rendering of the full-result page.
"FULL_RESULT_JSON"	JSON output of the full-result.
"MAPVIEW"	Search results displayed at a map (only test version).

Table C 1 Europeana Language Logger (ELL) actions

D. LIST OF FREQUENT CRAWLERS

"http://www.google.com/bot.html"
"http://help.yahoo.com/help/us/ysearch/slurp"
"http://yandex.com/bots"
"http://search.msn.com/msnbot.htm"
"http://www.pingdom.com/"
"nagios-plugins"
"http://www.dotnetdotcom.org/"
"http://www.majestic12.co.uk/bot.php"
"Yandex"
"Keybot Translation-Search-Machine"
"http://www.cuil.com/twiceler/robot.html"
"http://www.bing.com/bingbot.htm"
"http://www.baidu.com/search/spider.htm"
"http://www.scoutjet.com/"
"OpenSearchServer_Bot"
"FeedFetcher-Google-CoOp"
"http://yacy.net/bot.html"
"iCjobs"
"Europeana/1.0 (Europeana Test and Monitoring Client)"
"http://www.puritysearch.net/"
"http://www.google.com/feedfetcher.html"
"http://www.entireweb.com/about/search_tech/speedy_spider/"
"http://webagent.wise-guys.nl/"
"http://www.sitebot.org/robot/"
"http://yacy.net/bot.html"
"http://www.exabot.com/go/robot"

E. RESULTS FOR PAIR-WISE COUNTRY COMPARISONS

Country Pair	AD/CD	PL. RU	2.76	GR. IT	6.60	IT. NL	10.23
BR. PT	0.01	CH. DE	2.81	IE. RU	6.71	CA. HU	10.24
AT. CH	0.02	AT. HU	3.07	BE. FR	6.76	RO. SE	10.33
CH. PL	0.20	FR. HU	3.09	ES. NL	6.86	BR. DE	10.36
AT. PL	0.22	CH. HU	3.14	NL. RO	6.88	GB. SE	10.55
GB. IE	0.40	NO. RU	3.16	AT. US	6.94	GB. PL	10.57
BE. IT	0.46	CA. US	3.21	AT. BR	6.98	DE. PT	10.75
ES. NO	0.59	ES. IT	3.33	GR. RO	7.07	IE. NO	10.76
HU. IT	0.63	ES. RU	3.40	NO. PL	7.09	ES. FR	11.08
HU. RU	0.70	BE. BR	3.64	BR. CH	7.11	ES. PL	11.32
IE. US	0.72	BR. HU	3.76	GB. RU	7.16	BE. DE	11.41
ES. SE	0.83	BE. PT	3.76	AT. PT	7.17	SE. US	11.45
BE. SE	0.89	HU. PT	3.85	BR. RO	7.25	DE. US	11.54
BE. HU	0.90	GR. SE	3.85	CH. US	7.26	FR. GR	11.62
AT. FR	0.95	HU. PL	3.87	HU. NL	7.26	GB. NO	11.65
CH. FR	1.02	NL. NO	3.99	AT. ES	7.30	BR. IE	11.76
BR. GR	1.09	BR. IT	4.09	CH. PT	7.31	HU. RO	11.85
GR. PT	1.10	ES. GR	4.12	PT. RO	7.40	GR. PL	12.06
NO. SE	1.12	IT. PT	4.24	IE. PL	7.43	IE. PT	12.12
IT. SE	1.25	BR. RU	4.31	BE. PL	7.51	RO. RU	12.18
GR. NL	1.27	PT. RU	4.41	DE. GB	7.55	CA. SE	12.31
BR. NO	1.27	CA. GB	4.60	CH. ES	7.59	BE. IE	12.34
NO. PT	1.31	AT. SE	4.70	FR. IT	7.71	ES. RO	12.39
HU. SE	1.47	DE. RU	4.72	BR. FR	7.76	AT. NL	12.42
IT. RU	1.48	CH. SE	4.81	NL. RU	7.77	NO. US	12.52
BE. RU	1.69	DE. IE	4.81	HU. IE	7.91	BR. GB	12.62
FR. RU	1.97	DE. PL	4.88	RU. US	7.93	CH. NL	12.85
GB. US	2.02	AT. BE	5.03	AT. CA	7.99	GB. PT	13.09
CA. IE	2.05	AT. IT	5.04	DE. SE	8.03	IE. IT	13.31
RU. SE	2.12	FR. SE	5.13	FR. PT	8.06	CA. NO	13.35
BR. ES	2.13	BE. CH	5.21	IT. PL	8.29	DE. IT	13.37
AT. RU	2.19	CH. IT	5.25	BR. PL	8.33	BR. US	13.45
BE. NO	2.21	AT. IE	5.32	CA. CH	8.37	PL. US	13.62
ES. PT	2.21	GR. HU	5.42	GB. HU	8.54	BE. RO	13.79
CH. RU	2.24	CH. IE	5.48	PL. PT	8.63	PT. US	13.95
BR. NL	2.27	NL. SE	5.50	CA. RU	8.71	BR. CA	14.22
NL. PT	2.32	AT. NO	5.81	BE. NL	8.81	CA. PT	14.76
BR. SE	2.38	PL. SE	5.83	NO. RO	8.93	DE. GR	14.88
FR. PL	2.39	BE. GR	5.85	DE. NO	9.23	IT. RO	14.98
PT. SE	2.44	AT. GB	5.90	HU. US	9.39	BE. GB	15.21
BE. ES	2.50	CH. NO	5.94	AT. GR	9.44	CA. DE	15.42
HU. NO	2.55	GR. RU	5.97	CH. GR	9.68	ES. IE	15.61
GR. NO	2.57	DE. HU	5.99	IE. SE	9.71	GR. IE	15.65
IT. NO	2.63	CH. GB	6.15	FR. IE	9.82	DE. ES	16.14
ES. HU	2.66	FR. NO	6.45	DE. FR	10.09	CA. PL	16.28
AT. DE	2.71					AT. RO	16.49

FR. GB	16.73	**GR. US**	18.87	**ES. US**	22.68	**RO. US**	27.31
CH. RO	16.86	**CA. GR**	19.93	**CA. IT**	22.83	**NL. US**	28.24
BE. US	17.15	**ES. GB**	20.08	**IE. RO**	23.22	**CA. RO**	28.37
GB. GR	17.62	**FR. RO**	20.47	**DE. RO**	23.48	**CA. FR**	29.18
GB. IT	17.75	**PL. RO**	20.51	**FR. US**	23.53	**CA. NL**	29.98
NL. PL	17.89	**IT. US**	20.62	**CA. ES**	24.65		
FR. NL	18.19	**IE. NL**	20.98	**GB. NL**	25.84		
BE. CA	18.77	**DE. NL**	22.52	**GB. RO**	25.95		

Table E 1 Results for pair-wise country comparison: Browser locale

Country Pair	AD/CD	PT.SE	0.54	GB.RO	1.30	AT.GB	1.89
AT.FR	0.01	ES.RU	0.55	IT.SE	1.33	BE.RO	1.89
GR.PT	0.03	GR.SE	0.58	NL.RO	1.33	GR.PL	1.90
HU.RO	0.03	HU.RU	0.59	IT.PL	1.34	AT.NL	1.91
HU.US	0.05	RO.RU	0.61	CH.NL	1.36	AT.SE	1.94
GB.NL	0.07	ES.US	0.63	CH.GB	1.36	FR.HU	1.95
RO.US	0.10	PT.US	0.68	DE.IT	1.38	FR.GR	2.04
CH.IE	0.12	CH.FR	0.74	BE.FR	1.39	BR.IT	2.04
BE.CH	0.14	AT.BE	0.75	BE.PT	1.39	FR.IT	2.07
DE.IE	0.17	GB.PT	0.76	CH.PT	1.40	CA.PT	2.07
DE.PL	0.18	BR.ES	0.78	IE.PT	1.41	AT.RO	2.12
AT.PL	0.18	AT.CA	0.78	GR.IT	1.42	DE.RO	2.13
GR.RU	0.19	GR.US	0.80	CA.CH	1.44	BR.CH	2.14
PT.RU	0.20	BE.DE	0.80	IE.SE	1.44	BR.IE	2.14
RU.SE	0.24	NL.PT	0.80	CA.PL	1.44	PL.RO	2.14
HU.NO	0.24	DE.FR	0.83	BR.SE	1.45	BE.BR	2.15
IE.PL	0.25	BE.PL	0.84	CH.SE	1.49	BE.CA	2.15
BR.NO	0.25	GB.GR	0.84	BE.GB	1.52	PL.SE	2.15
BE.IE	0.25	IT.RU	0.85	CA.FR	1.52	NO.SE	2.16
ES.RO	0.27	BR.PT	0.86	BE.NL	1.54	DE.SE	2.16
ES.HU	0.27	RU.US	0.86	HU.IT	1.55	CA.HU	2.19
AT.DE	0.29	GR.NL	0.89	GR.IE	1.56	FR.RO	2.25
NO.RO	0.32	NO.PT	0.90	CH.GR	1.57	GB.PL	2.27
NO.US	0.32	BR.GR	0.93	CA.RU	1.57	ES.GB	2.30
ES.PT	0.33	HU.SE	0.96	BE.GR	1.58	DE.GB	2.34
AT.IE	0.33	BE.RU	0.97	BE.SE	1.58	BR.DE	2.34
GB.SE	0.36	ES.NO	0.98	ES.SE	1.59	AT.BR	2.34
BR.HU	0.38	BR.RU	0.99	DE.PT	1.64	BR.PL	2.36
FR.PL	0.39	CH.RU	0.99	BR.GB	1.65	NL.PL	2.36
CH.DE	0.40	IE.RU	1.02	BE.HU	1.65	CA.GR	2.37
ES.GR	0.42	GR.NO	1.04	AT.PT	1.65	GB.US	2.40
GB.RU	0.42	CA.IE	1.04	CH.HU	1.66	BR.FR	2.45
NL.SE	0.43	NO.RU	1.05	PL.PT	1.66	DE.NL	2.47
BE.IT	0.43	RO.SE	1.07	HU.IE	1.67	ES.NL	2.47
FR.IE	0.44	AT.IT	1.08	BR.NL	1.68	FR.SE	2.49
HU.PT	0.44	GB.HU	1.15	IT.RO	1.77	NL.US	2.52
BR.US	0.44	DE.RU	1.18	FR.PT	1.77	CA.RO	2.52
BR.RO	0.44	HU.NL	1.18	DE.HU	1.85	BR.CA	2.67
PT.RO	0.45	GB.IT	1.19	CA.DE	1.86	CA.IT	2.68
NL.RU	0.45	IT.NL	1.20	AT.GR	1.86	GB.NO	2.69
CH.IT	0.46	PL.RU	1.20	HU.PL	1.87	NL.NO	2.80
CH.PL	0.47	AT.RU	1.22	AT.HU	1.87	FR.GB	2.83
GR.HU	0.48	IT.PT	1.26	SE.US	1.87	CA.SE	2.96
GR.RO	0.50	IE.NL	1.28	DE.GR	1.88	FR.NL	3.03
AT.CH	0.52	GB.IE	1.29	IE.RO	1.88	ES.IE	3.14
IE.IT	0.52	FR.RU	1.29	CH.RO	1.88	IE.US	3.15

CA.GB	3.31	**BE.US**	3.58	**BE.ES**	4.01	**CA.US**	4.78
CH.US	3.33	**IT.NO**	3.63	**PL.US**	4.10	**ES.PL**	4.80
IT.US	3.38	**AT.US**	3.72	**DE.US**	4.16	**CA.NO**	4.96
IE.NO	3.39	**ES.IT**	3.75	**NO.PL**	4.31	**DE.ES**	5.03
CH.ES	3.47	**BE.NO**	3.82	**DE.NO**	4.37	**ES.FR**	5.58
CA.NL	3.48	**AT.NO**	3.94	**FR.US**	4.47	**CA.ES**	5.74
CH.NO	3.57	**AT.ES**	3.95	**FR.NO**	4.65		

Table E 2 Results for pair-wise country comparison: Google Language

Country Pair	AD/CD	DE.FR	2.78	BE.ES	5.50	BR.CA	8.27
GB.IE	0.13	CA.US	2.81	AT.NL	5.58	DE.HU	8.39
CA.CH	0.18	IE.US	2.83	CA.GB	5.60	HU.PL	8.54
IT.RU	0.34	AT.IT	2.84	CA.RO	5.76	CA.ES	8.55
AT.BE	0.38	GR.RU	2.90	CH.RO	5.76	DE.PT	8.63
NL.US	0.42	CA.PL	3.03	AT.US	5.77	FR.GR	8.68
BE.DE	0.44	CH.PL	3.05	HU.IT	5.81	IE.RO	8.73
IE.NO	0.50	AT.RO	3.07	AT.SE	5.87	IE.PL	8.74
BR.PT	0.52	NL.NO	3.11	NL.RU	5.96	PL.PT	8.75
GB.NO	0.54	ES.GR	3.12	RU.US	6.11	CH.GR	8.78
BR.GR	0.60	CA.SE	3.12	DE.IT	6.24	NO.RO	8.78
AT.DE	0.69	AT.CA	3.14	RU.SE	6.25	DE.SE	8.84
SE.US	0.82	IE.NL	3.15	IT.PL	6.28	CA.GR	8.90
ES.RU	0.86	BR.RU	3.17	BE.BR	6.31	FR.IT	8.93
DE.PL	0.86	AT.CH	3.19	AT.PT	6.39	BE.IE	9.07
ES.RO	0.88	FR.RU	3.23	BE.GR	6.69	FR.NL	9.10
ES.IT	1.03	BR.ES	3.32	BR.DE	6.84	FR.HU	9.23
BE.PL	1.09	AT.ES	3.48	AT.HU	6.89	FR.IE	9.28
AT.PL	1.13	HU.RO	3.55	AT.IE	7.01	GB.RO	9.34
NL.SE	1.16	BE.CH	3.71	BR.PL	7.02	FR.US	9.43
FR.PL	1.22	BE.CA	3.73	ES.PL	7.08	CH.HU	9.56
GR.PT	1.23	BE.RO	3.76	DE.ES	7.09	NO.PL	9.59
RO.RU	1.33	GR.IT	3.81	AT.NO	7.11	CA.HU	9.62
GR.RO	1.39	CH.DE	3.84	NO.RU	7.16	ES.FR	9.66
IT.RO	1.41	BR.IT	3.90	IE.RU	7.16	BE.NO	9.79
NO.SE	1.53	CA.DE	3.90	PL.SE	7.32	CH.PT	9.81
AT.RU	1.57	PT.RU	3.91	DE.GR	7.44	FR.PT	9.81
HU.PT	1.57	GB.US	4.11	GR.PL	7.59	CA.PT	9.94
BR.RO	1.80	CH.NO	4.12	CH.IT	7.62	DE.NL	9.97
IE.SE	1.82	DE.RO	4.13	NL.RO	7.64	DE.IE	10.18
AT.FR	1.84	CA.RU	4.17	NL.PL	7.67	DE.US	10.26
BR.HU	1.93	CH.IE	4.17	GB.RU	7.67	BR.NL	10.48
BE.RU	2.07	CH.RU	4.20	BE.SE	7.74	BR.SE	10.56
CH.NL	2.14	PL.RO	4.38	RO.US	7.79	BR.US	10.62
GB.SE	2.28	ES.PT	4.48	BR.FR	7.80	FR.NO	10.84
BE.FR	2.36	CA.IE	4.60	FR.SE	7.81	BR.IE	11.40
DE.RU	2.38	CA.NO	4.62	RO.SE	7.86	HU.NL	11.55
CH.US	2.39	GB.NL	4.63	BE.PT	7.87	BR.NO	11.58
PT.RO	2.42	BE.IT	4.69	BE.HU	7.91	DE.NO	11.60
CA.NL	2.55	HU.RU	4.83	CA.IT	7.91	HU.SE	11.64
GR.HU	2.63	CH.GB	4.94	AT.GB	7.94	HU.US	11.67
CA.FR	2.63	FR.RO	5.00	PL.US	7.96	GR.SE	11.74
CH.FR	2.65	IT.PT	5.16	BE.NL	7.99	GR.NL	11.84
PL.RU	2.66	AT.BR	5.27	BR.CH	8.21	GR.US	12.00
NO.US	2.72	AT.GR	5.29	CH.ES	8.25	BE.GB	12.02
CH.SE	2.74	ES.HU	5.30	BE.US	8.25	BR.GB	12.29

GB.PL	12.31	**PT.US**	12.94	**IT.NL**	14.01	**IT.NO**	15.17
HU.IE	12.38	**GR.NO**	13.03	**GB.GR**	14.12	**ES.NO**	15.60
HU.NO	12.52	**GB.HU**	13.10	**ES.IE**	14.15	**DE.GB**	15.94
GR.IE	12.64	**ES.SE**	13.20	**IT.US**	14.22	**FR.GB**	16.50
PT.SE	12.67	**IE.PT**	13.52	**ES.NL**	14.47	**GB.IT**	18.71
IT.SE	12.67	**IE.IT**	13.67	**ES.US**	14.68	**ES.GB**	18.90
NL.PT	12.79	**NO.PT**	13.90	**GB.PT**	14.93		

Table E 3 Results for pair-wise country comparison: Europeana Interface Language Change

Country pair	AD/CD	BR.RU	2.81	BE.HU	6.16	IT.SE	12.98
BE.DE	0.08	IE.US	2.88	FR.GR	6.27	ES.NL	13.88
AT.RU	0.19	CH.RU	2.91	CH.ES	6.32	IT.NO	13.91
BR.GR	0.25	GB.US	2.91	BR.CH	6.40	FR.NL	14.12
NL.SE	0.25	BE.ES	2.95	HU.PL	6.43	IT.NL	15.67
IT.RO	0.27	PT.RO	2.96	BE.PT	6.44	HU.US	32.34
AT.BE	0.30	CH.SE	3.00	DE.HU	6.47	HU.IE	32.68
DE.PL	0.31	CH.NL	3.17	AT.NL	6.51	GB.HU	33.39
BE.PL	0.32	AT.CH	3.21	AT.NO	6.64	CA.HU	33.75
AT.DE	0.37	ES.GR	3.26	FR.IT	6.75	RO.US	40.94
CA.IE	0.42	GR.RU	3.29	RO.SE	6.75	BR.US	40.99
ES.RO	0.43	AT.BR	3.35	PL.PT	6.85	BR.IE	41.03
BE.RU	0.46	HU.RO	3.45	FR.HU	6.90	IE.RO	41.05
DE.RU	0.52	ES.PL	3.53	DE.PT	6.97	GB.RO	42.24
AT.PL	0.53	CH.NO	3.55	NL.RO	6.99	BR.GB	42.39
FR.PL	0.53	DE.ES	3.70	NO.RO	7.15	CA.RO	42.60
NO.SE	0.63	CH.PL	3.74	CH.IT	7.39	BR.CA	42.77
PL.RU	0.65	BE.CH	3.81	CH.GR	7.54	RU.US	43.22
GB.IE	0.78	IT.PT	3.86	FR.PT	7.60	IE.RU	43.27
AT.FR	0.86	CH.FR	3.87	CH.HU	8.06	IE.PT	43.74
BE.FR	0.86	AT.GR	4.00	BE.SE	8.28	PT.US	43.92
FR.RU	0.90	HU.IT	4.12	PL.SE	8.58	GB.RU	44.59
HU.PT	0.93	BE.BR	4.15	CH.PT	8.65	CA.RU	44.95
NL.NO	1.02	CH.DE	4.23	BE.NO	9.07	GB.PT	45.61
DE.FR	1.13	BE.IT	4.25	BR.SE	9.09	CA.PT	46.03
ES.RU	1.19	PT.RU	4.33	NO.PL	9.45	GR.IE	47.90
RO.RU	1.24	CA.US	4.33	BR.NL	9.49	GR.US	48.33
GR.PT	1.27	CH.RO	4.36	BR.NO	9.52	GB.GR	50.23
ES.IT	1.34	BR.PL	4.44	BE.NL	9.68	CA.GR	50.66
BR.PT	1.39	BR.DE	4.44	DE.SE	9.89	AT.IE	54.30
BR.RO	1.53	HU.RU	4.63	FR.SE	10.00	AT.US	55.08
AT.RO	1.56	ES.PT	4.66	NL.PL	10.37	AT.GB	57.18
CA.GB	1.70	ES.HU	4.74	HU.SE	10.45	AT.CA	57.57
AT.ES	1.74	ES.FR	4.91	HU.NL	10.79	CH.IE	69.25
BR.IT	1.77	BR.FR	4.93	HU.NO	10.82	CH.US	71.62
GR.RO	1.89	IT.PL	4.96	GR.SE	10.88	CH.GB	74.93
IT.RU	1.91	AT.PT	5.15	DE.NO	10.92	CA.CH	75.26
GR.HU	2.02	BE.GR	5.15	FR.NO	11.18	BE.IE	82.46
BE.RO	2.04	RU.SE	5.23	GR.NO	11.40	IE.IT	83.22
BR.HU	2.09	AT.HU	5.27	ES.SE	11.58	ES.IE	83.54
DE.RO	2.19	DE.IT	5.33	GR.NL	11.61	BE.US	88.50
PL.RO	2.26	NL.RU	5.39	PT.SE	11.90	IT.US	90.53
GR.IT	2.41	GR.PL	5.54	NO.PT	12.40	ES.US	90.67
BR.ES	2.50	NO.RU	5.62	ES.NO	12.48	IE.PL	90.93
FR.RO	2.60	DE.GR	5.63	NL.PT	12.62	IE.SE	91.98
AT.IT	2.62	AT.SE	6.14	DE.NL	12.88	BE.GB	94.56

BE.CA	94.76	**PL.US**	99.78	**CA.NO**	109.95	**DE.GB**	137.07
IE.NO	94.91	**NO.US**	102.77	**GB.NO**	109.99	**FR.US**	150.33
ES.GB	97.62	**GB.SE**	105.68	**IE.NL**	117.24	**CA.NL**	150.57
CA.ES	97.70	**CA.SE**	105.69	**FR.IE**	121.70	**GB.NL**	151.89
GB.IT	97.71	**DE.IE**	105.82	**DE.US**	122.35	**CA.FR**	174.18
CA.IT	97.79	**CA.PL**	108.03	**NL.US**	135.87	**FR.GB**	177.54
SE.US	98.99	**GB.PL**	108.06	**CA.DE**	136.00		

Table E 4 Results for pair-wise country comparison: Usage of Native Interface Language

Country pair	AD/CD	IE.PL	0.74	IE.IT	1.90	NL.RO	2.79
DE.IT	0.03	BR.IE	0.75	BR.PL	1.91	ES.US	2.85
GR.RU	0.08	CH.NL	0.78	GB.IE	1.92	IT.SE	2.89
RO.RU	0.09	IT.US	0.81	GB.GR	1.93	CH.RO	2.90
CA.CH	0.10	AT.IT	0.82	GR.IT	1.94	CA.RU	2.95
GR.PT	0.11	AT.DE	0.84	FR.NL	1.94	IE.NL	2.95
IE.RO	0.13	GR.HU	0.84	AT.PT	1.97	HU.US	2.96
DE.GB	0.14	DE.US	0.86	RU.US	1.97	CH.PT	2.98
GB.IT	0.15	BR.RO	0.89	DE.IE	1.99	BE.RO	2.98
FR.NO	0.16	BE.NO	0.91	AT.RO	2.00	NL.PT	2.99
CA.FR	0.16	HU.PT	0.92	CH.IT	2.02	DE.SE	3.00
ES.PL	0.16	AT.CH	0.95	GB.PL	2.03	CH.IE	3.03
PT.RU	0.17	BR.RU	0.96	DE.NL	2.04	NO.RU	3.09
GR.RO	0.18	NL.US	0.96	DE.GR	2.07	AT.BR	3.12
ES.PT	0.21	CA.NL	0.97	AT.PL	2.08	CA.RO	3.13
IE.RU	0.21	AT.CA	1.10	CH.DE	2.09	CH.GR	3.13
CA.NO	0.23	NL.NO	1.20	CA.GB	2.11	FR.US	3.14
PT.RO	0.27	BR.GR	1.23	AT.GR	2.11	BE.IE	3.15
NO.SE	0.27	BE.SE	1.25	IT.PL	2.13	BR.GB	3.18
AT.US	0.27	ES.HU	1.26	AT.IE	2.14	BE.PT	3.21
CH.FR	0.28	AT.NO	1.29	RO.US	2.16	GR.NL	3.21
BR.HU	0.28	BE.US	1.30	ES.GB	2.16	BR.IT	3.25
PL.PT	0.32	BR.PT	1.31	AT.ES	2.17	NO.RO	3.26
CH.NO	0.32	BE.FR	1.35	BE.IT	2.19	CA.PT	3.26
GR.IE	0.32	HU.PL	1.36	PT.US	2.22	CA.IE	3.27
ES.GR	0.36	CH.US	1.43	SE.US	2.23	CH.PL	3.28
ES.RU	0.36	NL.SE	1.55	ES.IT	2.28	RU.SE	3.35
BE.NL	0.37	GB.NL	1.55	GB.NO	2.29	CH.ES	3.36
AT.NL	0.39	AT.SE	1.56	IE.US	2.32	NO.PT	3.40
IE.PT	0.41	IT.RU	1.56	CA.IT	2.33	IE.NO	3.40
PL.RU	0.44	AT.FR	1.58	BE.DE	2.37	BR.DE	3.42
HU.IE	0.45	GB.RU	1.59	GR.US	2.43	BE.GR	3.43
GR.PL	0.47	DE.RU	1.62	CA.DE	2.43	CA.GR	3.44
ES.RO	0.49	CA.US	1.67	DE.PL	2.44	RO.SE	3.53
BE.CH	0.50	IT.PT	1.74	IT.NO	2.52	CH.HU	3.56
CA.SE	0.51	IT.RO	1.74	GB.HU	2.54	GR.NO	3.57
FR.SE	0.53	GB.PT	1.74	HU.IT	2.56	HU.NL	3.61
PL.RO	0.57	GB.RO	1.76	NL.RU	2.58	BR.US	3.65
HU.RO	0.58	BR.ES	1.80	DE.ES	2.60	IE.SE	3.67
CH.SE	0.58	DE.RO	1.81	DE.NO	2.61	CA.PL	3.70
GB.US	0.59	CH.GB	1.84	AT.HU	2.65	PT.SE	3.71
AT.BE	0.65	DE.PT	1.85	GB.SE	2.65	CA.ES	3.79
HU.RU	0.65	BE.GB	1.86	DE.HU	2.67	BE.HU	3.80
ES.IE	0.65	AT.RU	1.86	PL.US	2.72	FR.RU	3.81
BE.CA	0.67	IT.NL	1.88	CH.RU	2.74	NL.PL	3.82
AT.GB	0.69	NO.US	1.88	BE.RU	2.78	NO.PL	3.82

CA.HU	3.83	**BR.CH**	4.09	**BR.NL**	4.38	**FR.HU**	5.06
FR.GB	3.85	**FR.RO**	4.10	**BR.CA**	4.43	**FR.GR**	5.09
GR.SE	3.89	**BE.ES**	4.17	**BR.NO**	4.53	**DE.FR**	5.85
ES.NO	3.91	**PL.SE**	4.21	**BE.BR**	4.56	**BR.FR**	6.18
HU.NO	3.94	**HU.SE**	4.23	**FR.PT**	4.74	**FR.PL**	7.55
ES.NL	3.95	**FR.IE**	4.28	**FR.IT**	4.81	**ES.FR**	7.68
BE.PL	4.05	**ES.SE**	4.30	**BR.SE**	4.84		

Table E 5 Results for pair-wise country comparison: Bounce Rate

Country Pair	AD/CD	AT.US	0.59	GR.PL	1.30	GR.SE	2.02
CH.IE	0.01	HU.RU	0.60	DE.NO	1.30	CH.DE	2.04
CH.FR	0.01	PL.PT	0.61	GR.US	1.31	GB.PT	2.05
FR.IE	0.02	GR.HU	0.63	CH.SE	1.31	BR.NO	2.07
BR.RO	0.05	PT.US	0.63	HU.IT	1.31	NO.RU	2.10
GR.RU	0.06	AT.PL	0.66	BE.CH	1.37	BE.ES	2.10
PL.US	0.07	NL.RU	0.67	AT.IE	1.39	NO.PT	2.11
IT.PT	0.10	BE.NO	0.69	DE.RO	1.40	NO.PL	2.13
CA.IE	0.13	PL.RO	0.73	RO.SE	1.41	BE.NL	2.18
CA.CH	0.14	BR.PL	0.73	ES.SE	1.41	IE.US	2.20
ES.NL	0.16	NO.SE	0.74	PT.SE	1.42	IE.RO	2.21
AT.DE	0.16	BE.DE	0.74	HU.NL	1.44	GB.PL	2.22
RO.RU	0.16	RO.US	0.75	DE.PL	1.44	DE.GR	2.22
BR.PT	0.17	ES.RU	0.75	BR.SE	1.46	HU.SE	2.25
BE.SE	0.19	BR.US	0.76	BR.DE	1.48	BR.IE	2.30
CA.FR	0.19	HU.RO	0.78	DE.PT	1.48	BE.FR	2.30
PT.RO	0.21	GR.IT	0.79	AT.CA	1.49	IE.PL	2.30
BR.RU	0.22	FR.NO	0.79	NL.SE	1.50	IE.RU	2.32
GR.RO	0.25	IT.PL	0.80	ES.HU	1.53	IE.PT	2.32
BR.IT	0.29	IT.US	0.82	RU.SE	1.55	ES.NO	2.34
NL.PT	0.30	AT.GB	0.83	DE.RU	1.55	DE.HU	2.34
ES.PL	0.30	AT.ES	0.85	AT.GR	1.56	CA.RO	2.36
IT.NL	0.31	BR.HU	0.86	AT.CH	1.56	DE.IT	2.40
DE.SE	0.32	GB.IE	0.89	DE.IE	1.63	CH.RO	2.40
IT.RO	0.32	PL.RU	0.90	BE.US	1.65	NL.NO	2.41
BR.GR	0.32	RU.US	0.92	BE.RO	1.67	CA.RU	2.46
ES.US	0.35	AT.NL	0.95	HU.PL	1.68	ES.IE	2.47
AT.SE	0.35	GR.NL	0.97	HU.US	1.69	ES.GB	2.48
BE.GB	0.37	CA.GB	0.97	FR.SE	1.77	BE.GR	2.48
PT.RU	0.39	AT.PT	1.00	BE.BR	1.77	BR.CA	2.49
GB.NO	0.40	AT.NO	1.04	DE.ES	1.77	CH.RU	2.51
ES.PT	0.41	AT.RO	1.06	IT.SE	1.79	BR.CH	2.53
NL.PL	0.45	HU.PT	1.07	BE.PT	1.80	IE.NL	2.54
BR.NL	0.46	CH.GB	1.08	FR.GB	1.80	GB.NL	2.54
CA.NO	0.46	AT.BR	1.08	BE.RU	1.81	BE.HU	2.57
GB.SE	0.47	ES.GR	1.09	BE.PL	1.83	CA.PT	2.59
NL.RO	0.48	SE.US	1.10	DE.NL	1.86	BE.IT	2.62
ES.IT	0.48	BE.IE	1.13	GB.RO	1.87	CH.PT	2.62
NL.US	0.49	IE.SE	1.14	AT.HU	1.89	CA.US	2.64
IE.NO	0.51	DE.GB	1.18	NO.RO	1.98	CH.US	2.64
IT.RU	0.52	AT.IT	1.18	BR.GB	1.99	GR.NO	2.71
GR.PT	0.55	PL.SE	1.20	NO.US	1.99	IT.NO	2.73
BR.ES	0.56	AT.RU	1.21	AT.FR	2.00	GB.GR	2.73
ES.RO	0.57	CA.SE	1.22	GB.RU	2.00	GB.HU	2.75
AT.BE	0.58	DE.US	1.24	GB.US	2.01	IE.IT	2.79
CH.NO	0.58	BE.CA	1.28	CA.DE	2.02	FR.RO	2.79

HU.NO	2.80	**CH.ES**	3.01	**CA.GR**	3.22	**FR.US**	4.34
CH.PL	2.81	**BR.FR**	3.03	**CH.GR**	3.23	**DE.FR**	4.51
CA.PL	2.82	**CA.ES**	3.04	**FR.PT**	3.32	**FR.PL**	4.97
GR.IE	2.86	**CH.NL**	3.06	**CH.IT**	3.41	**FR.NL**	5.11
FR.RU	2.88	**CA.NL**	3.09	**CA.IT**	3.47	**ES.FR**	5.24
HU.IE	2.96	**CA.HU**	3.17	**FR.HU**	3.66	**FR.IT**	6.09
GB.IT	3.00	**CH.HU**	3.20	**FR.GR**	4.10		

Table E 6 Results for pair-wise country comparison: External Access Point

Country Pair	AD/CD	BE.IT	1.19	NL.PL	2.48	BR.GB	3.68
CH.PL	0.01	IE.NO	1.20	DE.ES	2.49	HU.US	3.71
CA.IE	0.08	FR.IE	1.23	AT.SE	2.55	NL.RO	3.73
CH.SE	0.11	BR.HU	1.24	AT.RO	2.57	DE.PL	3.78
AT.BE	0.13	AT.NL	1.27	ES.RO	2.57	FR.RU	3.80
PL.SE	0.15	PT.RU	1.30	GB.RU	2.62	PT.US	3.84
FR.NO	0.24	GB.PL	1.30	BE.GB	2.63	IE.RO	3.87
PT.RO	0.25	BE.RU	1.31	CA.RU	2.63	CH.ES	3.88
CA.GB	0.26	NO.PL	1.32	ES.IE	2.64	NO.US	3.93
GR.HU	0.27	BE.DE	1.35	BR.US	2.68	BR.CH	3.99
GB.IE	0.29	DE.IE	1.35	ES.NL	2.69	AT.GR	4.00
ES.IT	0.35	RO.RU	1.36	SE.US	2.74	DE.HU	4.00
CH.IE	0.39	IE.US	1.53	IT.PT	2.74	BE.PL	4.06
HU.RO	0.43	CA.NO	1.64	AT.PL	2.74	CA.RO	4.11
IE.PL	0.44	ES.US	1.65	DE.SE	2.74	GB.IT	4.12
AT.ES	0.45	BR.IT	1.67	AT.PT	2.78	HU.NL	4.13
IE.SE	0.49	BE.NL	1.69	HU.IT	2.89	BR.SE	4.14
DE.US	0.49	RU.US	1.69	IE.IT	2.90	GB.RO	4.15
CA.CH	0.58	CA.DE	1.71	ES.PT	2.92	GR.IT	4.18
DE.NL	0.63	AT.IE	1.75	BE.RO	2.93	DE.NO	4.20
BR.RU	0.64	BR.GR	1.76	CH.RU	2.97	CH.IT	4.23
AT.US	0.65	CH.NL	1.76	AT.HU	2.99	HU.IE	4.24
BR.PT	0.65	HU.RU	1.77	DE.IT	2.99	ES.SE	4.25
AT.IT	0.68	BR.ES	1.84	ES.HU	3.02	DE.PT	4.30
CA.SE	0.72	DE.GB	1.84	BR.DE	3.02	ES.GR	4.34
HU.PT	0.72	CA.US	1.85	FR.GB	3.04	BR.PL	4.34
CA.PL	0.73	CA.FR	1.88	BE.CH	3.04	IE.PT	4.35
IT.RU	0.74	AT.BR	1.89	RU.SE	3.09	CH.RO	4.42
BE.US	0.74	FR.PL	1.90	IT.NL	3.11	NL.PT	4.42
GR.RO	0.76	AT.GB	1.92	BR.NL	3.19	FR.NL	4.48
BR.RO	0.80	GB.US	1.92	PL.RU	3.20	CA.HU	4.49
BE.ES	0.84	AT.CA	1.94	NL.NO	3.22	BE.NO	4.53
ES.RU	0.90	DE.RU	1.97	CA.ES	3.29	GB.HU	4.54
NO.SE	0.92	NL.SE	2.01	RO.US	3.29	RO.SE	4.54
CH.GB	0.93	BE.IE	2.02	AT.NO	3.32	IT.SE	4.64
NL.US	0.94	IT.US	2.02	BE.SE	3.34	PL.RO	4.70
IE.NL	0.95	NL.RU	2.16	PL.US	3.35	BE.GR	4.75
CH.NO	0.98	GB.NO	2.25	BR.IE	3.35	BR.NO	4.76
FR.SE	0.99	BE.BR	2.26	BE.PT	3.36	CH.HU	4.78
AT.DE	0.99	GR.RU	2.35	BE.HU	3.36	CA.PT	4.81
CH.FR	1.04	AT.CH	2.39	DE.RO	3.59	HU.SE	4.91
GB.NL	1.04	CH.DE	2.40	BR.CA	3.63	GB.PT	4.96
AT.RU	1.08	IT.RO	2.43	CA.IT	3.64	BR.FR	5.03
CA.NL	1.11	BE.CA	2.45	NO.RU	3.65	HU.PL	5.07
GB.SE	1.12	CH.US	2.46	AT.FR	3.65	NO.RO	5.07
GR.PT	1.18	IE.RU	2.47	ES.GB	3.67	CH.PT	5.18

GR.US	5.22	**ES.NO**	5.56	**IT.NO**	6.00	**GR.PL**	7.12
FR.US	5.27	**FR.HU**	5.61	**NO.PT**	6.10	**DE.FR**	7.24
FR.RO	5.27	**DE.GR**	5.73	**CA.GR**	6.10	**GR.NO**	7.40
ES.PL	5.33	**PL.PT**	5.74	**GB.GR**	6.33	**ES.FR**	7.63
PT.SE	5.38	**GR.NL**	5.81	**CH.GR**	6.46	**FR.GR**	7.96
HU.NO	5.42	**IT.PL**	5.89	**FR.PT**	6.58	**FR.IT**	8.42
GR.IE	5.54	**BE.FR**	5.94	**GR.SE**	6.68		

Table E 7 Results for pair-wise country comparison: Login

Country Pair	AD/CD	BR.RU	1.56	AT.NO	2.81	CA.RO	4.52
IE.NO	0.05	BR.PT	1.57	BR.IT	2.82	AT.PT	4.53
PL.SE	0.06	AT.DE	1.58	IE.RU	2.85	HU.NO	4.58
GR.PT	0.10	BR.GR	1.69	BE.RO	2.89	GB.HU	4.59
ES.RU	0.13	HU.PT	1.71	BE.CA	2.95	ES.PT	4.63
BR.HU	0.23	BE.RU	1.72	BE.ES	3.01	BR.IE	4.71
CA.IE	0.23	DE.IE	1.73	GB.RU	3.02	AT.GR	4.73
BE.US	0.32	ES.HU	1.78	NO.RU	3.12	IE.NL	4.75
AT.IT	0.33	AT.RO	1.78	HU.US	3.14	GB.IT	4.81
CA.NO	0.34	AT.CH	1.78	PT.RU	3.14	FR.GB	4.83
HU.RO	0.38	IT.US	1.81	AT.CA	3.18	CA.IT	4.88
GB.IE	0.46	GR.HU	1.83	CA.US	3.23	CA.HU	4.89
GB.NO	0.52	IT.RO	1.83	GR.RU	3.28	ES.GR	4.90
NL.RO	0.53	AT.NL	1.92	BE.HU	3.31	ES.NO	4.92
BR.RO	0.63	FR.IE	1.95	CH.RO	3.33	RO.SE	5.13
NL.RU	0.64	GB.SE	1.97	DE.RO	3.35	BE.PL	5.17
IT.RU	0.64	BE.IE	1.98	IE.IT	3.37	FR.RU	5.21
CH.GB	0.68	DE.RU	2.13	CH.ES	3.37	BR.NO	5.23
BE.DE	0.68	BE.IT	2.14	CA.RU	3.41	IT.PT	5.29
CH.DE	0.75	CH.SE	2.16	DE.IT	3.43	BR.GB	5.34
AT.RU	0.77	PT.RO	2.16	NL.PT	3.63	HU.SE	5.49
AT.US	0.84	AT.HU	2.18	CH.HU	3.72	PL.RO	5.53
CA.SE	0.87	IE.US	2.19	BR.US	3.75	CA.ES	5.56
CH.IE	0.90	DE.NO	2.22	CH.FR	3.77	PL.US	5.56
RO.RU	0.90	FR.PL	2.24	DE.HU	3.78	BR.CA	5.58
IE.SE	0.92	DE.GB	2.25	IE.RO	3.83	GR.IT	5.59
CA.GB	0.95	BR.ES	2.26	GR.NL	3.87	ES.GB	5.60
AT.ES	0.96	CH.RU	2.28	ES.IE	3.92	AT.FR	5.70
IE.PL	0.99	GR.RO	2.28	BE.BR	3.93	DE.NL	5.78
CA.PL	1.00	HU.IT	2.29	AT.SE	3.93	DE.PL	5.88
HU.NL	1.00	CA.FR	2.42	NL.US	3.97	HU.PL	5.90
ES.IT	1.01	BE.GB	2.42	BE.SE	3.98	NL.NO	5.92
CH.NO	1.01	BE.NO	2.44	DE.SE	3.99	IT.SE	6.01
AT.BE	1.06	IT.NL	2.45	RU.SE	4.02	PT.US	6.24
DE.US	1.08	AT.IE	2.47	HU.IE	4.18	BR.SE	6.25
BE.CH	1.11	CH.PL	2.47	GB.RO	4.20	BE.PT	6.43
NO.SE	1.18	GB.PL	2.55	NO.RO	4.21	CH.PT	6.45
HU.RU	1.27	AT.BR	2.59	IT.NO	4.26	FR.RO	6.46
FR.SE	1.28	ES.US	2.69	BE.NL	4.27	GR.US	6.55
ES.RO	1.32	NO.US	2.71	BR.CH	4.27	CA.NL	6.56
CA.CH	1.36	RO.US	2.72	SE.US	4.27	ES.SE	6.66
CH.US	1.36	CH.IT	2.72	PL.RU	4.33	IE.PT	6.70
NO.PL	1.36	FR.NO	2.73	CH.NL	4.35	CH.GR	6.71
BR.NL	1.40	AT.GB	2.73	DE.ES	4.40	BE.GR	6.75
ES.NL	1.44	GB.US	2.76	AT.PL	4.44	GB.NL	6.75
RU.US	1.55	CA.DE	2.80	BR.DE	4.50	FR.HU	6.81

BR.PL	6.82	**BR.FR**	7.89	**GB.GR**	8.40	**DE.FR**	10.68
GR.IE	6.93	**GR.NO**	7.91	**PT.SE**	8.79	**FR.PT**	11.19
DE.PT	7.28	**GB.PT**	8.02	**ES.PL**	8.85	**FR.GR**	11.74
NO.PT	7.60	**CA.PT**	8.04	**GR.SE**	9.14	**FR.IT**	11.90
NL.SE	7.64	**IT.PL**	8.15	**PL.PT**	9.80	**ES.FR**	12.38
DE.GR	7.66	**FR.US**	8.27	**NL.PL**	9.89	**FR.NL**	13.18
BE.FR	7.77	**CA.GR**	8.37	**GR.PL**	10.25		

Table E 8 Results for pair-wise country comparison: Search Sessions

Country Pair	AD/CD	AT.RO	0.90	CA.RO	1.76	BE.FR	2.71
BR.PT	0.03	HU.RO	0.92	AT.FR	1.79	BR.DE	2.73
AT.DE	0.03	BE.NL	0.92	CH.DE	1.80	BE.BR	2.74
GB.IE	0.07	FR.PL	0.92	FR.NL	1.80	CH.HU	2.78
IT.RO	0.09	PL.US	0.99	ES.HU	1.83	GR.IT	2.82
FR.IE	0.12	IE.NL	1.01	FR.RU	1.83	ES.PT	2.83
CA.PL	0.16	GB.PL	1.02	CA.ES	1.84	FR.NO	2.87
CH.PL	0.16	CA.SE	1.02	CH.NO	1.84	HU.PL	2.87
AT.BE	0.18	RU.US	1.03	AT.HU	1.84	HU.IE	2.88
NL.US	0.19	AT.CA	1.04	AT.GB	1.85	AT.NO	2.91
BE.DE	0.26	HU.IT	1.07	GB.US	1.87	NO.PL	2.96
CA.CH	0.26	GR.HU	1.07	SE.US	1.87	BR.US	3.09
IE.SE	0.28	DE.RO	1.08	IT.PT	1.88	NO.RO	3.12
ES.RU	0.28	IE.US	1.12	PT.RU	1.89	DE.SE	3.15
GB.SE	0.31	NL.RU	1.13	GB.RU	1.90	FR.HU	3.18
AT.ES	0.32	PL.SE	1.14	GB.NO	1.90	BE.PT	3.19
CH.IE	0.32	DE.US	1.15	CH.RO	1.91	BR.NL	3.19
BR.HU	0.34	BE.RO	1.15	AT.SE	1.92	GB.HU	3.21
CH.FR	0.34	BE.CA	1.17	DE.IE	1.93	DE.PT	3.22
HU.PT	0.39	IE.NO	1.21	FR.US	1.95	IT.US	3.22
RO.RU	0.39	AT.CH	1.23	RU.SE	1.98	HU.SE	3.25
FR.GB	0.41	NO.SE	1.25	PL.RO	1.98	ES.SE	3.25
CA.NL	0.45	HU.RU	1.29	IE.RO	2.04	BR.CA	3.26
AT.RU	0.46	AT.PL	1.31	CH.ES	2.06	CA.IT	3.27
CH.GB	0.52	CA.RU	1.31	DE.HU	2.10	BR.CH	3.38
IE.PL	0.53	BR.RO	1.32	BE.HU	2.14	IE.IT	3.40
CA.IE	0.56	BE.CH	1.41	ES.IE	2.17	CH.IT	3.44
DE.RU	0.56	DE.NL	1.42	GR.RO	2.18	BR.IE	3.46
CA.US	0.59	AT.IE	1.43	CA.NO	2.24	AT.GR	3.47
IT.RU	0.61	CH.RU	1.46	FR.RO	2.32	ES.GB	3.49
DE.ES	0.62	PT.RO	1.46	AT.BR	2.36	IT.NL	3.50
BE.RU	0.65	PL.RU	1.50	GB.RO	2.37	NL.NO	3.50
FR.SE	0.67	AT.IT	1.50	BE.IT	2.40	DE.GB	3.54
AT.US	0.67	ES.US	1.51	BR.ES	2.42	BR.PL	3.55
CH.SE	0.71	RO.US	1.51	RO.SE	2.43	NO.US	3.57
BE.US	0.71	CA.DE	1.55	BE.SE	2.46	PT.US	3.61
CA.FR	0.73	BR.IT	1.60	HU.US	2.46	NL.PT	3.73
CH.NL	0.73	BE.IE	1.60	DE.PL	2.53	ES.GR	3.74
BE.ES	0.73	NL.RO	1.61	BE.GB	2.54	CA.PT	3.75
BR.GR	0.74	IE.RU	1.62	HU.NL	2.55	HU.NO	3.86
GR.PT	0.77	BR.RU	1.71	DE.IT	2.61	CH.PT	3.88
ES.RO	0.79	BE.PL	1.72	GR.RU	2.62	BR.FR	3.89
AT.NL	0.80	GB.NL	1.74	CA.HU	2.65	BR.GB	3.90
NL.PL	0.82	ES.NL	1.75	ES.PL	2.66	ES.FR	3.91
CH.US	0.86	NL.SE	1.75	AT.PT	2.66	BR.SE	3.91
CA.GB	0.87	ES.IT	1.75	NO.RU	2.68	IE.PT	3.92

BE.GR	4.08	**PT.SE**	4.53	**GR.IE**	4.72	**GB.GR**	5.43
DE.GR	4.15	**GB.PT**	4.56	**IT.SE**	4.79	**FR.GR**	5.50
BE.NO	4.15	**BR.NO**	4.59	**GR.PL**	5.04	**DE.NO**	5.56
PL.PT	4.16	**CA.GR**	4.60	**ES.NO**	5.16	**FR.IT**	5.95
DE.FR	4.44	**FR.PT**	4.60	**GB.IT**	5.25	**GR.NO**	6.21
IT.PL	4.49	**GR.NL**	4.62	**NO.PT**	5.36	**IT.NO**	6.84
GR.US	4.50	**CH.GR**	4.71	**GR.SE**	5.37		

Table E 9 Results for pair-wise country comparison: Browsing Sessions

Country Pair	p.value (holm)
AT . CA	0
AT . CH	0
AT . GR	0
AT . HU	0
AT . IT	0
AT . NL	0
AT . NO	0
AT . PL	0
AT . PT	0
AT . RU	0
AT . SE	0
AT . US	0
BE . CA	0
BE . CH	0
BE . DE	0
BE . FR	0
BE . GR	0
BE . HU	0
BE . IT	0
BE . NL	0
BE . NO	0
BE . PL	0
BE . PT	0
BE . RU	0
BE . SE	0
BE . US	0
BR . CA	0
BR . CH	0
BR . GB	0
BR . GR	0
BR . NL	0
BR . PL	0
BR . PT	0
BR . US	0
CA . DE	0
CA . ES	0
CA . FR	0
CA . GB	0
CA . GR	0
CA . HU	0
CA . IE	0
CA . IT	0
CA . NL	0
CA . NO	0
CA . PL	0
CA . PT	0
CA . RU	0
CA . SE	0
CH . DE	0
CH . ES	0
CH . FR	0
CH . GB	0
CH . GR	0
CH . HU	0
CH . IE	0
CH . IT	0
CH . NL	0
CH . NO	0
CH . PL	0
CH . PT	0
CH . RU	0
CH . SE	0
DE . ES	0
DE . GB	0
DE . GR	0
DE . IE	0
DE . NL	0
DE . NO	0
DE . PL	0
DE . PT	0
DE . RO	0
DE . SE	0
DE . US	0
ES . FR	0
ES . GR	0
ES . HU	0
ES . IT	0
ES . NL	0
ES . NO	0
ES . PL	0
ES . PT	0
ES . RU	0
ES . SE	0
ES . US	0
FR . GB	0
FR . GR	0
FR . HU	0
FR . IE	0
FR . NL	0
FR . NO	0
FR . PL	0
FR . PT	0
FR . RO	0
FR . SE	0
FR . US	0
GB . GR	0
GB . HU	0
GB . IT	0
GB . NL	0
GB . NO	0
GB . PL	0
GB . PT	0
GB . RU	0
GB . SE	0
GB . US	0
GR . HU	0
GR . IE	0
GR . IT	0
GR . NO	0
GR . PL	0
GR . RO	0
GR . RU	0
GR . SE	0
GR . US	0
HU . IE	0
HU . NL	0
HU . PL	0
HU . PT	0
HU . RO	0
HU . US	0
IE . IT	0
IE . NL	0
IE . NO	0
IE . PL	0
IE . PT	0
IE . RU	0
IE . SE	0
IE . US	0
IT . NL	0
IT . PL	0
IT . PT	0
IT . RO	0
IT . US	0
NL . NO	0
NL . PL	0
NL . RO	0
NL . RU	0
NL . SE	0
NL . US	0
NO . PL	0
NO . PT	0
NO . RO	0
NO . US	0
PL . PT	0
PL . RO	0
PL . RU	0
PL . SE	0
PL . US	0
PT . RO	0
PT . RU	0
PT . SE	0
PT . US	0
RO . RU	0
RO . SE	0
RU . US	0
SE . US	0
AT . FR	0.05
AT . RO	0.05
BE . RO	0.05
BR . NO	0.05
BR . RO	0.05
IT . NO	0.05
AT . GB	0.10
BE . GB	0.10
DE . HU	0.10
FR . RU	0.10
DE . RU	0.13
BR . ES	0.17
BR . HU	0.21
BR . IE	0.21
BE . ES	0.24
BR . SE	0.27
AT . DE	0.30
BR . RU	0.30
NL . PT	0.58
AT . ES	0.60
IT . SE	0.64
FR . IT	0.76
AT . IE	0.77
GR . NL	0.77
HU . IT	0.87
AT . BE	1
AT . BR	1

BE . BR	1	**CA . US**	1	**ES . RO**	1	**IE . RO**	1
BE . IE	1	**CH . RO**	1	**GB . IE**	1	**IT . RU**	1
BR . DE	1	**CH . US**	1	**GB . RO**	1	**NO . RU**	1
BR . FR	1	**DE . FR**	1	**GR . PT**	1	**NO . SE**	1
BR . IT	1	**DE . IT**	1	**HU . NO**	1	**RO . US**	1
CA . CH	1	**ES . GB**	1	**HU . RU**	1	**RU . SE**	1
CA . RO	1	**ES . IE**	1	**HU . SE**	1		

Table E 10 Results for pair-wise country comparison: Unique queries per Session

Country Pair	p.value (holm)
AT . BR	0
AT . CA	0
AT . CH	0
AT . ES	0
AT . FR	0
AT . GB	0
AT . GR	0
AT . HU	0
AT . IE	0
AT . IT	0
AT . NO	0
AT . PL	0
AT . PT	0
AT . RO	0
AT . RU	0
AT . SE	0
AT . US	0
BE . BR	0
BE . CA	0
BE . CH	0
BE . ES	0
BE . FR	0
BE . GB	0
BE . GR	0
BE . HU	0
BE . IE	0
BE . IT	0
BE . NO	0
BE . PL	0
BE . PT	0
BE . RO	0
BE . RU	0
BE . SE	0
BR . CA	0
BR . CH	0
BR . DE	0
BR . ES	0
BR . FR	0
BR . GB	0
BR . GR	0
BR . HU	0
BR . IE	0
BR . IT	0
BR . NL	0
BR . NO	0
BR . PL	0
BR . PT	0
BR . RO	0
BR . RU	0
BR . SE	0
BR . US	0
CA . DE	0
CA . ES	0
CA . FR	0
CA . GB	0
CA . GR	0
CA . HU	0
CA . IE	0
CA . IT	0
CA . NL	0
CA . PL	0
CA . PT	0
CA . RO	0
CA . RU	0
CA . US	0
CH . DE	0
CH . ES	0
CH . FR	0
CH . GR	0
CH . HU	0
CH . IE	0
CH . IT	0
CH . NL	0
CH . PL	0
CH . PT	0
CH . RO	0
CH . RU	0
CH . US	0
DE . ES	0
DE . FR	0
DE . GB	0
DE . GR	0
DE . HU	0
DE . IE	0
DE . IT	0
DE . NO	0
DE . PL	0
DE . PT	0
DE . RO	0
DE . RU	0
DE . SE	0
DE . US	0
ES . FR	0
ES . GB	0
ES . GR	0
ES . HU	0
ES . IE	0
ES . IT	0
ES . NL	0
ES . NO	0
ES . RO	0
ES . RU	0
ES . SE	0
ES . US	0
FR . GB	0
FR . GR	0
FR . HU	0
FR . IT	0
FR . NL	0
FR . NO	0
FR . PL	0
FR . PT	0
FR . RO	0
FR . RU	0
FR . SE	0
FR . US	0
GB . GR	0
GB . HU	0
GB . IE	0
GB . IT	0
GB . NL	0
GB . NO	0
GB . PL	0
GB . PT	0
GB . RO	0
GB . RU	0
GB . SE	0
GB . US	0
GR . IE	0
GR . IT	0
GR . NL	0
GR . NO	0
GR . PL	0
GR . PT	0
GR . RO	0
GR . SE	0
GR . US	0
HU . IE	0
HU . IT	0
HU . NL	0
HU . NO	0
HU . PL	0
HU . PT	0
HU . RO	0
HU . SE	0
HU . US	0
IE . IT	0
IE . NL	0
IE . NO	0
IE . PL	0
IE . PT	0
IE . RO	0
IE . RU	0
IE . SE	0
IT . NL	0
IT . NO	0
IT . PL	0
IT . PT	0
IT . RU	0
IT . SE	0
IT . US	0
NL . NO	0
NL . PL	0
NL . PT	0
NL . RO	0
NL . RU	0
NL . SE	0
NO . PL	0
NO . PT	0
NO . RO	0
NO . RU	0
NO . US	0
PL . RO	0
PL . RU	0
PL . SE	0
PL . US	0
PT . RO	0
PT . RU	0
PT . SE	0
PT . US	0
RO . RU	0
RO . SE	0
RO . US	0

RU . SE	0	**IE . US**	0.16	**BE . NL**	1	**ES . PT**	1
RU . US	0	**DE . NL**	0.22	**CA . CH**	1	**GR . HU**	1
SE . US	0	**FR . IE**	0.25	**CA . NO**	1	**GR . RU**	1
NL . US	0.02	**HU . RU**	0.31	**CA . SE**	1	**IT . RO**	1
BE . US	0.07	**AT . BE**	1	**CH . NO**	1	**NO . SE**	1
CH . GB	0.07	**AT . DE**	1	**CH . SE**	1	**PL . PT**	1
BE . DE	0.15	**AT . NL**	1	**ES . PL**	1		

Table E 11 Results for pair-wise country comparison: Duration in Minutes

Country Pair	AD/CD	PL.RU	1.98	BE.PL	3.87	BR.US	6.20
PT.RU	0.02	ES.GR	2.03	BE.RU	3.90	CA.RU	6.25
CA.CH	0.04	AT.US	2.06	NL.RO	3.92	ES.US	6.29
ES.RO	0.07	BR.RO	2.10	AT.CA	3.94	NO.US	6.31
BE.DE	0.11	AT.ES	2.14	DE.NL	3.98	PT.US	6.39
PL.RO	0.11	NO.SE	2.14	DE.RU	4.03	AT.NO	6.39
BR.HU	0.22	HU.RO	2.19	BE.CH	4.04	FR.RU	6.45
GR.RU	0.25	BE.RO	2.21	AT.FR	4.07	HU.NL	6.47
GB.SE	0.27	GR.PL	2.26	BE.ES	4.16	BE.GB	6.64
GR.PT	0.34	DE.RO	2.27	BE.CA	4.37	CH.IT	6.65
ES.PL	0.34	GB.NO	2.30	AT.SE	4.49	IT.NL	6.65
BR.RU	0.50	FR.IE	2.35	CH.DE	4.60	RU.SE	6.68
BR.PT	0.56	ES.PT	2.37	DE.PL	4.62	IE.RU	6.75
IT.RO	0.60	IT.RU	2.45	CH.RO	4.63	NO.RO	6.77
CA.SE	0.65	DE.US	2.47	AT.IE	4.66	BR.NL	6.88
CH.SE	0.65	CH.US	2.57	CA.RO	4.79	CH.HU	6.97
GB.IE	0.66	PL.PT	2.60	GB.US	4.79	GR.NL	7.03
HU.RU	0.68	CH.NO	2.67	BE.GR	4.80	CA.HU	7.16
AT.DE	0.72	NL.SE	2.71	BE.HU	4.84	NL.PT	7.20
AT.BE	0.72	BR.ES	2.72	FR.RO	4.87	CA.IT	7.22
HU.PT	0.75	ES.HU	2.73	RU.US	4.92	GB.RU	7.24
IE.SE	0.78	CA.US	2.81	DE.ES	4.95	CH.PL	7.25
CH.FR	0.80	CA.NO	2.81	DE.HU	5.01	BR.CH	7.32
BR.GR	0.87	AT.NL	2.85	BE.BR	5.05	IE.IT	7.36
CA.FR	0.94	GR.IT	2.90	BE.PT	5.05	CH.GR	7.39
IT.PL	0.95	HU.PL	2.91	BE.IE	5.05	NL.PL	7.41
IE.NO	0.98	BR.PL	2.92	CA.DE	5.08	FR.HU	7.46
CH.GB	0.99	AT.RU	2.97	FR.NO	5.10	CH.ES	7.47
CA.GB	1.03	IE.NL	3.05	BE.SE	5.13	CH.PT	7.55
GR.HU	1.04	BE.IT	3.06	DE.GR	5.14	BR.CA	7.57
NL.US	1.15	FR.GB	3.09	AT.GB	5.20	HU.SE	7.60
ES.IT	1.29	BE.NL	3.11	IT.US	5.24	HU.IE	7.60
CH.IE	1.30	FR.US	3.19	RO.SE	5.24	ES.NL	7.68
AT.IT	1.33	IT.PT	3.21	BR.DE	5.27	CA.GR	7.72
CA.IE	1.33	RO.US	3.29	DE.PT	5.37	CA.PL	7.85
AT.RO	1.41	AT.GR	3.34	IE.RO	5.40	CA.PT	7.86
GR.RO	1.46	HU.IT	3.38	NL.NO	5.47	BE.NO	7.89
RO.RU	1.47	BR.IT	3.45	BE.FR	5.49	IE.PL	7.89
PT.RO	1.72	SE.US	3.57	NL.RU	5.51	BR.IE	7.96
CH.NL	1.75	AT.PT	3.60	DE.IE	5.58	GR.IE	8.04
ES.RU	1.81	DE.IT	3.67	GB.RO	5.78	IT.SE	8.04
FR.SE	1.84	AT.CH	3.75	HU.US	5.87	BR.SE	8.05
FR.NL	1.85	IE.US	3.76	DE.SE	5.97	BR.FR	8.05
AT.PL	1.92	GB.NL	3.78	PL.US	6.02	CA.ES	8.08
BE.US	1.93	AT.HU	3.78	CH.RU	6.08	ES.IE	8.09
CA.NL	1.93	AT.BR	3.82	GR.US	6.19	NO.RU	8.13

IE.PT	8.19	**PL.SE**	8.66	**GB.PT**	9.46	**FR.PL**	11.18
GB.HU	8.22	**FR.PT**	8.69	**BR.NO**	9.70	**IT.NO**	11.20
GR.SE	8.28	**BR.GB**	8.83	**GR.NO**	10.29	**ES.GB**	11.40
DE.FR	8.29	**ES.SE**	8.88	**NO.PT**	10.35	**ES.FR**	11.42
PT.SE	8.41	**HU.NO**	9.07	**FR.IT**	10.47	**NO.PL**	11.78
DE.GB	8.44	**DE.NO**	9.35	**GB.IT**	10.53	**ES.NO**	11.99
FR.GR	8.61	**GB.GR**	9.39	**GB.PL**	11.18		

Table E 12 Results for pair-wise country comparison: Brief Result Paging

Country pair	AD/CD	GB.IT	0.60	CA.GR	1.43	DE.NL	2.27
IE.NL	0.01	FR.PL	0.61	CA.DE	1.45	CA.PL	2.28
AT.CH	0.02	GR.RU	0.63	DE.PL	1.46	ES.US	2.30
HU.RU	0.04	BE.CH	0.65	CH.GR	1.46	PT.US	2.44
FR.SE	0.05	AT.BE	0.65	BR.IE	1.48	DE.RO	2.45
GR.RO	0.05	BE.DE	0.69	PT.RO	1.51	AT.FR	2.46
AT.CA	0.09	GR.HU	0.69	ES.RO	1.53	BR.ES	2.46
CA.IE	0.09	BE.IE	0.74	IT.SE	1.55	IE.PT	2.48
CA.CH	0.11	BE.ES	0.74	GR.NL	1.59	GB.US	2.49
CA.NL	0.12	ES.HU	0.76	DE.HU	1.61	BE.GR	2.50
CH.ES	0.14	ES.RU	0.80	DE.RU	1.63	CH.FR	2.52
AT.ES	0.16	BE.CA	0.84	IT.NL	1.64	NL.SE	2.62
AT.IE	0.17	AT.IT	0.85	BE.SE	1.65	BR.GB	2.63
CH.IE	0.18	CH.IT	0.85	FR.NO	1.67	FR.RU	2.67
RO.US	0.20	IE.RO	0.89	NL.US	1.68	FR.HU	2.68
AT.NL	0.22	GB.NL	0.90	DE.ES	1.69	ES.PL	2.74
BE.IT	0.24	IE.IT	0.90	GB.RO	1.72	BE.FR	2.76
CH.NL	0.25	IE.US	0.94	IT.PL	1.72	GR.IT	2.82
CA.ES	0.30	GB.HU	0.96	GR.PT	1.75	IE.NO	2.83
HU.IE	0.32	BR.US	0.97	BE.PL	1.80	DE.NO	2.84
GR.US	0.33	BR.PT	0.99	IE.PL	1.81	BE.US	2.85
BE.GB	0.33	GB.RU	0.99	IE.SE	1.83	BE.BR	2.87
ES.IE	0.33	GR.IE	1.06	NO.PL	1.84	DE.FR	2.88
IE.RU	0.36	DE.GB	1.07	AT.BR	1.86	FR.IT	2.89
PL.SE	0.38	CA.IT	1.08	AT.SE	1.89	RO.SE	2.90
ES.GB	0.38	BR.RU	1.08	GB.SE	1.89	CA.FR	2.95
CH.GB	0.40	ES.IT	1.08	BR.CA	1.89	IT.NO	2.99
AT.GB	0.41	BR.HU	1.14	BR.CH	1.90	AT.NO	3.02
HU.NL	0.43	CA.RO	1.14	CH.SE	1.90	BE.NO	3.02
RU.US	0.46	DE.IE	1.16	BE.RO	1.93	CH.NO	3.05
CA.HU	0.47	AT.RO	1.16	AT.PL	1.93	AT.PT	3.06
NL.RU	0.48	BE.HU	1.16	CH.PL	1.96	PL.RO	3.07
DE.IT	0.48	AT.DE	1.17	ES.GR	2.03	CH.PT	3.14
CA.RU	0.51	CH.DE	1.18	PT.RU	2.03	FR.GB	3.15
RO.RU	0.51	NL.RO	1.19	BR.NL	2.07	BR.IT	3.15
AT.HU	0.53	CH.RO	1.19	HU.PT	2.12	NO.RU	3.21
HU.US	0.53	BE.RU	1.20	CA.SE	2.12	CA.PT	3.21
GB.IE	0.54	NO.SE	1.22	GB.PL	2.13	HU.NO	3.21
CH.HU	0.55	BE.NL	1.26	IT.RO	2.14	NL.PL	3.22
BR.RO	0.56	DE.SE	1.33	FR.IE	2.20	GB.NO	3.25
AT.RU	0.56	HU.IT	1.34	HU.SE	2.21	DE.GR	3.29
HU.RO	0.57	AT.US	1.35	RU.SE	2.22	CA.NO	3.33
CA.GB	0.57	IT.RU	1.37	GB.GR	2.23	IT.US	3.38
ES.NL	0.58	CA.US	1.39	HU.PL	2.25	GR.SE	3.49
CH.RU	0.59	CH.US	1.41	ES.SE	2.26	BR.DE	3.55
BR.GR	0.59	AT.GR	1.42	PL.RU	2.26	FR.RO	3.58

ES.NO	3.66	**ES.FR**	4.14	**FR.NL**	4.67	**PT.SE**	5.23
NL.PT	3.75	**BR.PL**	4.17	**BR.NO**	4.83	**DE.PT**	5.70
BR.SE	3.76	**DE.US**	4.21	**FR.GR**	4.86	**PL.PT**	6.23
SE.US	3.80	**ES.PT**	4.24	**BR.FR**	4.87	**NO.PT**	6.36
NO.RO	3.92	**GB.PT**	4.33	**PL.US**	4.91	**FR.US**	6.76
GR.PL	3.99	**BE.PT**	4.62	**IT.PT**	5.08	**FR.PT**	7.44
NL.NO	3.99	**GR.NO**	4.65	**NO.US**	5.12		

Table E 13 Results for pair-wise country comparison: Full Result Paging

Country Pair	AD/CD	PT.RO	0.87	DE.GR	1.65	BR.US	2.38
AT.BE	0.03	BE.CA	0.88	GB.RU	1.66	BE.IE	2.39
IE.SE	0.05	RO.RU	0.88	AT.NL	1.66	BR.CA	2.40
ES.RU	0.06	GB.NO	0.90	IE.RU	1.70	FR.GR	2.40
AT.PL	0.15	CH.US	0.92	RO.US	1.71	NL.RO	2.44
PL.RU	0.17	AT.IT	0.94	GR.US	1.73	ES.PT	2.51
ES.PL	0.21	ES.GR	0.94	FR.GB	1.74	FR.PL	2.52
AT.RU	0.24	DE.PL	0.97	CA.GR	1.75	BE.NL	2.52
BE.PL	0.26	BR.GR	0.97	CA.RO	1.76	GB.RO	2.55
CA.US	0.26	ES.IT	0.98	RU.SE	1.77	IE.RO	2.55
HU.PT	0.27	AT.GR	0.99	FR.SE	1.77	FR.NO	2.61
AT.ES	0.29	HU.RO	1.00	AT.BR	1.79	PL.PT	2.63
BE.RU	0.30	FR.US	1.06	NL.US	1.80	RO.SE	2.63
IE.NO	0.30	GR.PL	1.08	BR.ES	1.81	DE.HU	2.64
CH.FR	0.31	PL.US	1.09	AT.IE	1.81	HU.US	2.68
DE.US	0.33	CA.NL	1.09	CH.PL	1.82	DE.NL	2.69
GR.IT	0.33	CA.PL	1.13	IE.US	1.82	CA.HU	2.70
GB.NL	0.34	ES.RO	1.14	AT.GB	1.82	NO.US	2.71
GB.IE	0.34	FR.RU	1.17	HU.IT	1.84	BE.PT	2.72
GB.SE	0.36	AT.CH	1.17	HU.RU	1.84	ES.FR	2.72
BR.PT	0.37	DE.ES	1.19	CA.NO	1.86	IE.PL	2.72
CH.NL	0.40	AT.FR	1.19	PT.RU	1.89	BR.CH	2.75
AT.DE	0.40	CH.RU	1.19	BR.PL	1.91	CH.IT	2.76
GR.RO	0.40	AT.RO	1.19	BE.FR	1.93	BE.GB	2.77
NO.SE	0.44	IT.PL	1.20	AT.SE	1.94	BE.SE	2.78
IT.RU	0.45	BE.GR	1.21	CH.ES	1.97	GR.NL	2.79
BE.ES	0.45	CH.NO	1.22	BE.BR	2.00	GR.IE	2.82
CA.FR	0.46	NL.NO	1.22	NO.RU	2.00	NO.RO	2.84
BR.RO	0.48	PL.RO	1.24	IT.PT	2.01	DE.SE	2.85
CA.DE	0.53	CH.DE	1.28	DE.FR	2.03	BR.FR	2.85
AT.US	0.56	ES.US	1.28	GB.US	2.07	ES.IE	2.86
CA.CH	0.56	CA.ES	1.29	CA.IT	2.08	GB.GR	2.94
IE.NL	0.57	CA.GB	1.30	CH.RO	2.10	CH.HU	3.00
BR.HU	0.58	CA.IE	1.32	SE.US	2.14	DE.GB	3.00
BE.DE	0.59	BE.RO	1.34	FR.RO	2.14	GR.SE	3.00
DE.RU	0.59	FR.NL	1.34	AT.HU	2.19	NL.PL	3.06
GR.RU	0.61	BR.IT	1.35	ES.HU	2.21	FR.HU	3.06
CH.GB	0.62	FR.IE	1.38	CH.GR	2.21	CA.PT	3.11
NL.SE	0.65	BE.IT	1.38	IT.US	2.25	DE.PT	3.12
AT.CA	0.69	BR.RU	1.40	AT.NO	2.26	BR.NL	3.14
IT.RO	0.70	CA.SE	1.46	DE.IE	2.28	PT.US	3.15
RU.US	0.71	GR.PT	1.48	HU.PL	2.29	BR.IE	3.20
BE.US	0.77	GR.HU	1.48	DE.IT	2.32	ES.NL	3.23
CH.IE	0.77	BE.CH	1.52	BR.DE	2.33	PL.SE	3.24
CA.RU	0.81	NL.RU	1.53	BE.HU	2.37	BR.GB	3.25
CH.SE	0.83	DE.RO	1.64	AT.PT	2.38	HU.NL	3.30

BR.SE	3.32	**HU.SE**	3.45	**FR.PT**	3.73	**PT.SE**	4.18
GB.PL	3.32	**ES.GB**	3.47	**NO.PL**	3.88	**IT.NL**	4.28
GR.NO	3.32	**CH.PT**	3.50	**FR.IT**	3.92	**IT.SE**	4.32
BE.NO	3.34	**BR.NO**	3.54	**IE.PT**	4.00	**NO.PT**	4.44
HU.IE	3.38	**HU.NO**	3.63	**NL.PT**	4.02	**GB.IT**	4.51
ES.SE	3.38	**IE.IT**	3.66	**ES.NO**	4.03	**IT.NO**	5.00
GB.HU	3.39	**DE.NO**	3.66	**GB.PT**	4.14		

Table E 14 Results for pair-wise country comparison: Selection of Language Facet

Country pair	AD/CD	IE.PL	0.58	BE.PL	1.14	HU.NL	1.68
AT.ES	0	AT.RO	0.59	ES.IE	1.15	BE.PT	1.71
HU.PT	0.02	AT.PL	0.60	BR.PT	1.15	GB.PL	1.71
CH.GB	0.03	CH.FR	0.60	BR.US	1.16	AT.NO	1.72
CA.GB	0.04	IE.NO	0.61	DE.GB	1.17	IT.NO	1.73
NL.RU	0.06	DE.SE	0.62	BR.IE	1.18	CA.RO	1.75
CA.CH	0.07	IE.RU	0.63	IT.RO	1.19	CH.RO	1.76
GR.HU	0.11	DE.RU	0.65	GB.US	1.19	BR.GB	1.78
IT.PL	0.12	ES.RO	0.68	NO.RU	1.24	NO.PL	1.80
CA.SE	0.14	AT.IT	0.68	DE.NL	1.25	ES.GR	1.80
GR.PT	0.15	RO.RU	0.69	CH.IT	1.26	GB.RO	1.85
DE.IE	0.17	HU.RO	0.70	CA.IT	1.26	HU.PL	1.87
AT.BE	0.17	GR.RO	0.71	BE.IT	1.27	ES.SE	1.88
AT.RU	0.18	BR.NL	0.71	BE.IE	1.28	DE.ES	1.89
CH.SE	0.20	BE.NL	0.72	AT.SE	1.30	BR.NO	1.90
GB.SE	0.20	DE.IT	0.73	RO.US	1.31	ES.PT	1.91
ES.RU	0.21	SE.US	0.75	AT.HU	1.32	HU.IT	1.92
IE.SE	0.23	CA.FR	0.75	IE.RO	1.33	NO.RO	1.96
BR.RO	0.24	NL.US	0.82	DE.NO	1.33	BE.DE	1.96
BE.ES	0.25	IE.NL	0.82	CH.PL	1.33	BE.SE	1.98
BE.BR	0.27	CA.DE	0.83	BE.HU	1.34	HU.IE	1.99
CH.NO	0.28	PT.RO	0.83	CA.PL	1.34	GB.NL	2.02
FR.NO	0.28	DE.PL	0.85	ES.US	1.35	HU.US	2.03
PL.RU	0.28	CH.DE	0.85	BR.DE	1.35	CH.ES	2.06
DE.US	0.29	AT.US	0.85	NO.US	1.36	NL.NO	2.08
IE.US	0.31	FR.IE	0.87	HU.RU	1.37	GR.NL	2.08
AT.NL	0.33	FR.SE	0.90	NL.SE	1.43	CA.ES	2.10
IT.US	0.33	RU.SE	0.91	AT.CA	1.43	AT.FR	2.14
CA.IE	0.34	AT.IE	0.91	AT.CH	1.45	BE.CH	2.15
IT.RU	0.34	NL.RO	0.92	DE.RO	1.47	NL.PT	2.17
BE.RU	0.35	CA.US	0.93	ES.HU	1.47	DE.HU	2.18
CA.NO	0.36	BR.PL	0.95	BE.US	1.49	BE.CA	2.19
AT.BR	0.36	CH.US	0.95	AT.GR	1.50	FR.RO	2.23
CH.IE	0.38	BR.HU	0.97	FR.RU	1.50	BR.FR	2.23
GB.NO	0.38	ES.PL	0.97	GR.RU	1.52	HU.SE	2.31
GB.IE	0.40	FR.GB	1.00	BR.SE	1.54	GR.IE	2.34
NL.PL	0.43	CA.RU	1.01	AT.GB	1.58	GR.PL	2.35
BR.ES	0.43	BR.IT	1.01	BE.GR	1.60	FR.US	2.39
PL.US	0.44	CH.RU	1.03	AT.PT	1.61	CA.HU	2.40
NO.SE	0.49	BR.GR	1.03	CH.NL	1.62	CH.HU	2.40
BR.RU	0.49	IT.SE	1.05	PT.RU	1.63	IE.PT	2.42
RU.US	0.50	AT.DE	1.07	CA.NL	1.63	PL.PT	2.43
ES.NL	0.51	GB.RU	1.09	GB.IT	1.64	GR.IT	2.43
IE.IT	0.51	ES.IT	1.11	RO.SE	1.65	GB.HU	2.50
BE.RO	0.53	PL.RO	1.13	BR.CA	1.65	IT.PT	2.50
IT.NL	0.55	PL.SE	1.13	BR.CH	1.66	GR.US	2.55

ES.NO	2.56	**FR.HU**	2.84	**CA.PT**	3.05	**NO.PT**	3.30
HU.NO	2.59	**DE.PT**	2.86	**DE.FR**	3.16	**FR.NL**	3.34
ES.GB	2.60	**GR.SE**	2.87	**FR.IT**	3.16	**FR.PT**	3.75
BE.NO	2.61	**PT.SE**	2.92	**FR.PL**	3.18	**FR.GR**	3.75
BE.GB	2.62	**CH.GR**	2.98	**GB.GR**	3.21	**BE.FR**	3.83
PT.US	2.62	**CA.GR**	3.00	**GB.PT**	3.24	**ES.FR**	4.08
DE.GR	2.81	**CH.PT**	3.03	**GR.NO**	3.26		

Table E 15 Results for pair-wise country comparison: Selection of Country Facet

Country Pair	AD/CD	IE.IT	0.44	PT.US	1.01	FR.HU	1.60
BR.GR	0.00	GB.PT	0.45	NO.PL	1.01	CH.SE	1.65
ES.HU	0.00	GB.NL	0.45	IT.NO	1.02	BE.PT	1.67
IT.PL	0.01	CA.FR	0.46	HU.US	1.02	BE.GB	1.70
CH.FR	0.04	DE.FR	0.48	CH.HU	1.03	NL.RU	1.72
NL.NO	0.04	NO.RO	0.51	NL.PT	1.05	DE.GB	1.75
BR.SE	0.05	CA.IE	0.51	CA.NO	1.06	BE.ES	1.78
GR.SE	0.05	GB.HU	0.52	PL.PT	1.06	BR.ES	1.79
BE.DE	0.06	IE.SE	0.53	ES.PL	1.07	BE.SE	1.79
IE.PT	0.08	AT.IE	0.56	HU.NL	1.07	DE.PT	1.80
AT.CA	0.10	CH.PL	0.56	ES.SE	1.08	CA.NL	1.80
BE.CH	0.10	BR.IE	0.57	AT.ES	1.09	DE.SE	1.81
HU.PT	0.13	GR.IE	0.57	AT.NO	1.11	ES.GR	1.83
AT.DE	0.13	CH.IT	0.57	IT.PT	1.12	CA.US	1.83
NO.US	0.13	BR.RO	0.58	AT.PT	1.12	FR.SE	1.93
GB.IE	0.13	NO.PT	0.58	ES.IT	1.14	AT.NL	1.94
HU.IE	0.13	GR.RO	0.59	RO.US	1.18	FR.GB	1.97
ES.IE	0.14	ES.GB	0.60	CH.ES	1.20	HU.RO	1.97
CH.DE	0.14	BR.US	0.60	CH.NO	1.20	CH.NL	1.98
AT.BE	0.15	GB.SE	0.61	CA.GB	1.22	AT.US	1.99
ES.PT	0.16	GR.US	0.61	CH.PT	1.23	DE.ES	2.01
NL.US	0.17	HU.NO	0.63	BE.NO	1.25	CH.US	2.01
BR.NO	0.18	DE.IE	0.64	DE.NO	1.25	PT.RO	2.03
GR.NO	0.18	BE.IE	0.64	GB.PL	1.26	NL.PL	2.03
NO.SE	0.18	CH.IE	0.65	IE.RU	1.26	GB.RU	2.04
AT.CH	0.20	ES.NO	0.66	ES.NL	1.28	RU.US	2.04
CA.PL	0.24	FR.IE	0.72	BE.HU	1.28	IT.NL	2.09
CA.IT	0.24	RO.RU	0.76	ES.US	1.29	FR.PT	2.15
CA.DE	0.25	HU.PL	0.76	GB.IT	1.29	PL.US	2.20
BE.CA	0.27	BE.PL	0.77	FR.PL	1.30	BR.CA	2.22
CA.CH	0.29	CA.HU	0.79	GB.RO	1.31	CA.GR	2.26
BE.FR	0.29	HU.IT	0.79	AT.GB	1.33	IT.US	2.29
NL.SE	0.30	BR.GB	0.82	DE.HU	1.34	ES.RO	2.31
GB.NO	0.31	BE.IT	0.82	FR.NO	1.35	AT.BR	2.37
IE.US	0.31	GB.GR	0.83	BR.RU	1.38	BR.CH	2.37
IE.NO	0.33	IE.RO	0.85	GR.RU	1.40	CH.GR	2.40
GB.US	0.34	DE.PL	0.85	CH.GB	1.42	AT.GR	2.41
RO.SE	0.34	RU.SE	0.89	FR.IT	1.46	ES.FR	2.50
AT.FR	0.35	AT.HU	0.91	BR.HU	1.49	BE.NL	2.51
AT.PL	0.37	NL.RO	0.93	CA.SE	1.50	BR.PL	2.58
AT.IT	0.38	DE.IT	0.94	PL.SE	1.50	GR.PL	2.65
IE.NL	0.38	CA.ES	0.94	BR.PT	1.51	DE.NL	2.65
BR.NL	0.39	PT.SE	0.97	GR.HU	1.51	CA.RO	2.65
GR.NL	0.39	NO.RU	0.97	IT.SE	1.52	BR.IT	2.66
SE.US	0.42	CA.PT	0.99	GR.PT	1.54	GR.IT	2.74
IE.PL	0.44	HU.SE	1.00	AT.SE	1.58	BE.US	2.77

HU.RU	2.77	**BE.BR**	3.07	**CA.RU**	3.42	**FR.RO**	3.90
CH.RO	2.77	**IT.RO**	3.10	**BE.RO**	3.47	**PL.RU**	3.95
AT.RO	2.79	**BE.GR**	3.15	**CH.RU**	3.50	**IT.RU**	4.05
PT.RU	2.89	**ES.RU**	3.23	**BR.FR**	3.55	**BE.RU**	4.38
FR.NL	2.92	**BR.DE**	3.25	**AT.RU**	3.58	**DE.RU**	4.59
DE.US	3.00	**DE.GR**	3.35	**DE.RO**	3.63	**FR.RU**	4.88
PL.RO	3.02	**FR.US**	3.38	**FR.GR**	3.67		

Table E 16 Results for pair-wise country comparison: Selection of Native Language Facet

Country Pair	AD/CD	PL.PT	0.39	GR.RO	0.80	GB.NO	1.45
DE.IT	0.00	CH.GB	0.40	FR.HU	0.80	ES.RO	1.46
AT.SE	0.00	AT.PT	0.43	IT.SE	0.80	NL.RU	1.49
BE.PT	0.02	FR.PL	0.44	ES.GR	0.82	CH.IT	1.50
NL.RO	0.02	IT.NO	0.44	DE.SE	0.82	CH.NO	1.53
IE.PT	0.02	ES.IT	0.44	IE.RO	0.82	CH.DE	1.53
GR.SE	0.03	AT.BE	0.45	CH.IE	0.82	PL.RO	1.54
BE.IE	0.03	DE.NO	0.46	AT.NL	0.84	GB.IT	1.56
AT.GR	0.03	BE.PL	0.46	HU.NO	0.85	CH.FR	1.62
CH.RO	0.08	GB.NL	0.46	BE.NO	0.85	NO.RO	1.63
ES.IE	0.11	GB.GR	0.48	CH.HU	0.89	DE.GB	1.63
CH.NL	0.11	DE.ES	0.49	FR.SE	0.90	NL.PT	1.66
HU.IE	0.12	HU.PL	0.49	GB.PT	0.93	GB.RU	1.67
FR.IT	0.15	GR.PT	0.49	IE.NL	0.93	IT.RO	1.71
BE.HU	0.16	BE.GR	0.53	HU.RO	0.93	RU.SE	1.72
HU.PT	0.16	IE.NO	0.54	CH.RU	0.93	DE.RO	1.76
ES.PL	0.18	ES.SE	0.57	GR.PL	0.94	FR.GB	1.77
DE.FR	0.18	NO.PL	0.58	AT.IT	0.94	IE.RU	1.77
IE.PL	0.18	GB.IE	0.60	BE.FR	0.96	BE.NL	1.86
HU.SE	0.22	IT.PT	0.61	NO.SE	0.96	FR.RO	1.87
ES.PT	0.25	RO.SE	0.61	AT.DE	0.97	AT.RU	1.90
AT.HU	0.25	CH.SE	0.62	BE.GB	0.99	NL.NO	1.97
IT.PL	0.25	DE.PT	0.64	AT.NO	1.05	HU.RU	2.21
GR.HU	0.26	PL.SE	0.65	AT.FR	1.08	GR.RU	2.22
DE.PL	0.27	GB.HU	0.66	CH.PT	1.09	ES.NL	2.27
IE.SE	0.29	HU.IT	0.67	GR.NL	1.11	NL.PL	2.36
IE.IT	0.30	ES.FR	0.67	BE.CH	1.12	PT.RU	2.62
DE.IE	0.30	AT.ES	0.67	GR.NO	1.14	IT.NL	2.65
BE.ES	0.31	AT.CH	0.67	RO.RU	1.16	NO.RU	2.77
AT.IE	0.31	AT.RO	0.67	GR.IT	1.17	BE.RU	2.78
GR.IE	0.31	ES.NO	0.69	PT.RO	1.18	DE.NL	2.87
GB.SE	0.34	DE.HU	0.69	HU.NL	1.19	ES.RU	3.05
GB.RO	0.36	BE.IT	0.72	DE.GR	1.23	PL.RU	3.11
FR.IE	0.36	NL.SE	0.73	BE.RO	1.24	FR.NL	3.12
AT.GB	0.37	CH.GR	0.77	ES.GB	1.26	IT.RU	3.30
PT.SE	0.38	AT.PL	0.77	CH.ES	1.30	DE.RU	3.41
ES.HU	0.38	BE.DE	0.78	GB.PL	1.36	FR.RU	3.55
BE.SE	0.38	FR.PT	0.78	CH.PL	1.37		
FR.NO	0.38	NO.PT	0.79	FR.GR	1.39		

Table E 17 Results for pair-wise country comparison: Selection of Native Country Facet

Country Pair	AD/CD	AT.SE	3.87	CH.IT	7.85	GB.PT	13.28
GR.PT	0.02	BE.CA	3.91	GR.NO	8.15	RO.SE	13.59
CH.GB	0.17	CH.NL	3.98	CA.CH	8.17	AT.GR	14.12
CA.RO	0.35	IE.PT	4.23	CH.US	8.24	HU.NO	14.24
HU.RU	0.35	PT.US	4.64	NL.PL	8.29	DE.NL	14.36
FR.PL	0.53	CH.ES	4.77	BR.CA	8.36	ES.SE	14.37
CH.PL	0.88	GR.IE	4.78	AT.BE	8.43	NO.RU	14.40
CH.NO	0.91	IT.NO	4.86	HU.RO	8.50	PL.PT	14.70
AT.FR	0.97	ES.IT	5.05	GR.HU	8.53	BR.IT	14.71
IT.US	1.01	NO.SE	5.17	ES.PT	8.62	BE.SE	14.83
GB.PL	1.11	ES.RO	5.19	BE.GB	8.68	PL.US	14.90
GB.NO	1.17	NO.US	5.26	RO.RU	8.69	CA.DE	15.14
CH.FR	1.17	CH.SE	5.34	GR.RU	8.79	IT.PL	15.21
AT.PL	1.21	GR.US	5.34	NL.PT	9.31	CA.SE	15.22
ES.NL	1.23	IT.PT	5.48	GB.RO	9.33	GB.GR	15.37
BR.IE	1.38	BR.PT	5.54	AT.RO	9.55	BE.BR	15.40
AT.CH	1.60	CA.ES	5.63	FR.NL	9.72	FR.PT	15.55
BR.HU	1.63	ES.US	5.65	ES.PL	9.82	CH.IE	15.76
FR.GB	1.66	NO.RO	5.66	ES.GR	10.06	BE.DE	15.92
CA.US	1.71	DE.PL	5.74	PL.RO	10.30	ES.IE	16.12
NO.PL	1.73	IE.RO	5.74	AT.CA	10.33	DE.ES	16.28
PT.RO	1.83	CA.NO	5.78	CA.HU	10.41	FR.US	16.82
RO.US	1.87	NL.RO	5.82	CA.GB	10.43	IE.NL	16.86
AT.GB	1.88	BE.CH	5.88	CH.PT	10.48	HU.US	17.23
AT.DE	1.92	DE.GB	5.99	BE.PL	10.62	GR.PL	17.33
BR.RU	1.93	IT.NL	6.17	CA.RU	10.63	BR.CH	17.34
GR.RO	1.95	DE.FR	6.27	FR.RO	10.82	RU.US	17.50
NL.NO	1.95	GB.NL	6.28	AT.IT	10.82	FR.IT	17.58
FR.NO	1.96	BR.GR	6.29	GR.NL	10.84	PT.SE	17.89
BE.ES	2.06	CA.NL	6.36	IE.US	11.02	DE.PT	18.19
AT.NO	2.25	GR.IT	6.39	AT.US	11.10	SE.US	18.37
BE.IT	2.35	AT.NL	6.50	CH.GR	11.44	IT.SE	18.59
CA.PT	2.42	NL.US	6.68	ES.FR	11.51	FR.GR	18.62
CA.IT	2.46	BE.PT	6.75	IE.NO	11.54	BR.ES	18.64
IT.RO	2.50	BR.RO	6.92	CA.PL	11.71	AT.IE	18.83
ES.NO	2.54	CA.IE	6.96	BE.FR	11.94	BR.NL	19.37
CA.GR	2.63	PL.SE	6.97	GB.IT	12.29	BE.HU	19.57
BE.US	3.09	FR.SE	7.12	GB.US	12.36	HU.IT	19.58
HU.IE	3.10	GB.SE	7.26	IE.IT	12.38	BE.RU	19.82
BE.NL	3.11	HU.PT	7.33	CA.FR	12.46	IT.RU	19.85
IE.RU	3.37	AT.ES	7.41	BR.NO	12.65	CH.HU	20.04
DE.SE	3.40	PT.RU	7.55	AT.PT	12.77	CH.RU	20.24
BE.NO	3.51	ES.GB	7.56	DE.RO	13.05	GR.SE	20.37
CH.DE	3.75	CH.RO	7.70	BR.US	13.13	AT.BR	20.67
BE.RO	3.77	NO.PT	7.70	NL.SE	13.13	DE.US	21.09
DE.NO	3.86	BE.GR	7.73	BE.IE	13.27	GB.IE	21.29

DE.GR	21.80	**ES.RU**	24.49	**DE.IE**	28.78	**HU.PL**	34.64
DE.IT	22.37	**HU.NL**	24.95	**BR.FR**	28.99	**PL.RU**	34.85
IE.PL	23.78	**NL.RU**	25.19	**GB.HU**	29.92	**FR.HU**	38.66
BR.GB	23.92	**FR.IE**	25.51	**GB.RU**	30.14	**FR.RU**	38.81
AT.HU	24.10	**IE.SE**	25.96	**BR.DE**	32.48	**DE.HU**	43.02
ES.HU	24.24	**BR.PL**	26.85	**HU.SE**	34.26	**DE.RU**	43.18
AT.RU	24.31	**BR.SE**	28.52	**RU.SE**	34.44		

Table E 18 Results for pair-wise country comparison: Selection of Native Language Collections

Country Pair	AD/CD	AT.PT	4.56	GB.RO	9.60	GR.SE	16.36
GB.NO	0.24	CH.PT	4.74	PL.RO	10.60	DE.PT	16.55
HU.PT	0.27	NO.SE	4.78	GB.GR	10.63	NO.RU	16.59
AT.CH	0.37	CH.RU	4.90	ES.HU	10.83	HU.SE	17.23
BE.RO	0.45	DE.GB	4.98	FR.NL	10.88	CH.NL	17.46
NO.PL	0.80	DE.SE	5.24	ES.IT	11.02	BE.PL	17.65
BE.GR	0.84	AT.RO	5.52	ES.PT	11.06	PT.SE	17.70
GR.RO	1.00	AT.RU	5.55	ES.SE	11.12	AT.NL	17.89
GB.PL	1.16	CH.RO	5.68	IE.RO	11.24	FR.IT	19.42
ES.NO	1.17	NL.RO	5.96	FR.RO	11.49	CH.IE	20.18
PT.RO	1.26	GR.NL	6.10	GR.IE	11.92	IT.SE	20.26
FR.NO	1.33	IT.NO	6.26	GR.PL	12.11	AT.IE	20.36
FR.PL	1.39	ES.IE	6.33	GB.HU	12.25	BE.FR	20.40
HU.RO	1.47	ES.FR	6.36	CH.IT	12.45	CH.ES	20.92
IE.SE	1.50	GR.NO	6.53	GB.PT	12.53	IT.RU	21.26
GR.IT	1.55	GB.NL	6.54	AT.IT	12.64	BE.SE	21.49
DE.IE	1.83	NO.RO	6.69	DE.RO	12.79	AT.ES	21.54
BE.PT	2.18	IT.NL	6.69	AT.NO	12.80	DE.IT	22.12
IT.RO	2.38	FR.SE	7.49	CH.NO	12.85	CH.GB	22.21
BE.HU	2.40	BE.NO	7.72	IE.IT	12.87	BE.DE	22.74
FR.GB	2.44	PL.SE	7.87	GB.IT	13.05	AT.GB	22.82
DE.NO	2.47	AT.GR	8.08	BE.ES	13.10	CH.PL	24.94
GR.PT	2.56	GB.SE	8.13	HU.IE	13.38	IE.RU	25.80
GR.HU	2.75	CH.GR	8.18	GR.RU	13.42	AT.PL	25.84
ES.GB	2.84	ES.RO	8.20	FR.GR	13.46	NL.RU	27.07
NL.NO	3.07	NO.PT	8.24	HU.PL	13.48	CH.FR	27.46
IE.NO	3.17	HU.NO	8.29	IE.PT	13.54	CH.SE	27.72
BE.IT	3.23	HU.RU	8.30	DE.NL	13.55	AT.SE	28.43
FR.IE	3.34	HU.NL	8.40	PL.PT	13.87	AT.FR	28.68
DE.FR	3.69	NL.PT	8.49	NL.SE	14.31	CH.DE	29.38
IE.PL	3.87	NL.PL	8.64	RO.SE	14.38	AT.DE	30.71
ES.NL	4.05	IE.NL	8.71	BE.IE	14.40	ES.RU	31.54
AT.HU	4.05	ES.GR	8.96	FR.HU	14.57	GB.RU	32.18
CH.HU	4.24	AT.BE	9.03	BE.GB	14.89	PL.RU	37.18
GB.IE	4.39	BE.CH	9.04	FR.PT	15.07	RU.SE	37.33
DE.PL	4.44	PT.RU	9.07	DE.GR	15.15	FR.RU	42.14
IT.PT	4.45	DE.ES	9.15	DE.HU	15.98	DE.RU	44.49
ES.PL	4.51	BE.NL	9.19	IT.PL	16.13		
HU.IT	4.56	RO.RU	9.55	BE.RU	16.26		

Table E 19 Results for pair-wise country comparison: Selection of Native Country Collections

Country Pair	AD/CD	NL.SE	1.40	CH.HU	2.67	RO.US	4.62
CA.US	0.01	AT.RU	1.45	CA.GB	2.68	BR.CA	4.71
BE.ES	0.02	DE.SE	1.46	HU.PT	2.70	CA.IE	4.74
CH.DE	0.09	ES.IT	1.51	GR.NL	2.75	NO.PT	4.75
HU.RU	0.16	BE.DE	1.55	CA.NO	2.78	BR.DE	4.75
AT.SE	0.18	AT.DE	1.55	BR.IT	2.81	FR.GB	4.76
BR.GR	0.19	HU.IT	1.65	ES.US	2.82	AT.GB	4.81
BR.RO	0.24	DE.US	1.65	PT.RU	2.83	AT.PL	4.82
CA.FR	0.26	NL.RO	1.66	BR.SE	2.88	FR.RO	4.83
IT.NO	0.27	AT.HU	1.66	BE.RO	2.98	GB.SE	5.00
GB.PL	0.27	RU.SE	1.67	DE.RU	3.02	NL.US	5.03
BE.SE	0.32	DE.ES	1.68	AT.GR	3.04	PL.SE	5.04
ES.SE	0.34	CH.NO	1.72	ES.RO	3.05	AT.PT	5.06
FR.US	0.38	NO.RO	1.77	BE.FR	3.09	FR.PL	5.12
AT.IT	0.42	CH.IT	1.86	GB.IE	3.15	IE.US	5.27
GR.RO	0.44	HU.SE	1.90	CH.RO	3.20	CH.IE	5.34
HU.RO	0.45	CA.SE	2.02	CA.IT	3.24	BR.US	5.34
NL.NO	0.45	DE.FR	2.03	NO.US	3.27	NL.PT	5.35
AT.BE	0.50	BE.CA	2.07	CA.RU	3.31	CA.GR	5.49
AT.ES	0.53	AT.CA	2.08	DE.HU	3.34	PT.SE	5.58
AT.NO	0.55	BE.RU	2.10	IE.PL	3.40	FR.IT	5.60
RO.RU	0.60	BR.PT	2.10	GB.US	3.40	DE.GB	5.65
IT.SE	0.70	BR.NL	2.13	PL.US	3.42	GB.RU	5.66
BR.HU	0.70	GR.PT	2.14	ES.FR	3.44	PL.RU	5.66
NO.SE	0.78	CA.ES	2.14	GR.SE	3.44	BR.FR	5.69
NL.RU	0.82	ES.RU	2.15	FR.NO	3.50	GB.NO	5.85
BE.CH	0.85	BR.NO	2.16	DE.IT	3.51	FR.IE	5.91
BR.RU	0.86	AT.RO	2.17	CH.PL	3.51	NO.PL	5.92
CH.ES	0.87	PT.RO	2.22	BE.BR	3.55	DE.GR	5.97
CA.CH	0.93	BE.NL	2.25	CH.GB	3.55	BE.GB	6.00
GR.HU	0.94	IT.RO	2.27	CA.HU	3.58	GB.HU	6.02
CH.FR	0.98	BE.HU	2.36	GR.IT	3.58	DE.PL	6.03
CH.SE	0.99	SE.US	2.38	BR.ES	3.66	HU.PL	6.03
NO.RU	1.03	AT.US	2.40	BR.CH	3.67	BE.PL	6.23
IT.NL	1.04	ES.HU	2.43	RU.US	3.67	AT.IE	6.26
HU.NL	1.06	ES.NL	2.43	FR.RU	3.78	IT.PT	6.27
AT.NL	1.07	CH.RU	2.44	CA.NL	3.87	CH.PT	6.32
CH.US	1.08	RO.SE	2.44	HU.US	3.98	FR.NL	6.33
GR.RU	1.11	DE.NO	2.45	DE.RO	4.01	ES.GB	6.37
AT.CH	1.11	AT.FR	2.46	FR.HU	4.12	IE.SE	6.39
CA.DE	1.18	CH.NL	2.48	CA.RO	4.14	DE.IE	6.47
BE.NO	1.23	FR.SE	2.49	CH.GR	4.27	GR.US	6.47
HU.NO	1.24	AT.BR	2.57	IT.US	4.31	GB.RO	6.69
ES.NO	1.29	CA.PL	2.60	BE.GR	4.38	ES.PL	6.69
BE.IT	1.38	GR.NO	2.63	DE.NL	4.39	PL.RO	6.73
IT.RU	1.39	BE.US	2.67	ES.GR	4.56	BE.IE	6.87

BE.PT	6.94	**BR.GB**	7.59	**BR.IE**	8.38	**GR.IE**	9.26
IE.RU	6.94	**CA.PT**	7.66	**IT.PL**	8.43	**GR.PL**	9.29
IE.NO	6.99	**BR.PL**	7.68	**GB.NL**	8.45	**FR.PT**	10.19
ES.IE	7.02	**IE.RO**	7.74	**DE.PT**	8.82	**IE.PT**	11.05
FR.GR	7.20	**IE.IT**	7.89	**NL.PL**	8.93	**GB.PT**	11.62
ES.PT	7.22	**GB.IT**	7.92	**PT.US**	9.05	**PL.PT**	11.98
HU.IE	7.22	**IE.NL**	8.31	**GB.GR**	9.06		

Table E 20 Results for pair-wise country comparison: Outlink to Content Provider